Universidade Técnica de Lisboa
Instituto Superior de Economia e Gestão

Mestrado em: Ciências Actuariais

A estimação de tarifas na presença de franquias e limites de indemnização
—
uma aplicação à cobertura de Choque, Colisão ou Capotamento do seguro automóvel

Rui Alexandre Silva Esteves

Orientação: Professor Doutor João Manuel de Sousa Andrade e Silva

Júri: **Presidente** Professor Doutor João Manuel de Sousa Andrade e Silva
Vogais Professor Doutor João Tiago Praça Nunes Mexia
Professor Doutor José Manuel Matos Passos

Fevereiro de 2002

Universidade Técnica de Lisboa
Instituto Superior de Economia e Gestão

Mestrado em: Ciências Actuariais

A estimação de tarifas na presença de franquias e limites de indemnização

—

uma aplicação à cobertura de Choque, Colisão ou Capotamento do seguro automóvel

Rui Alexandre Silva Esteves

Orientação: Professor Doutor João Manuel de Sousa Andrade e Silva

Júri: **Presidente** Professor Doutor João Manuel de Sousa Andrade e Silva
Vogais Professor Doutor João Tiago Praça Nunes Mexia
Professor Doutor José Manuel Matos Passos

Fevereiro de 2002

Resumo

A construção de estruturas tarifárias é, habitualmente, feita em dois passos: estimação da indemnização média por sinistro e estimação da frequência de sinistralidade. Os modelos lineares generalizados são utilizados de forma genérica na construção dessas estruturas, para seguros com um número considerável de unidades em risco e um reduzido grau de heterogeneidade entre elas, podendo ser utilizados na estimação de qualquer uma das parcelas.

Apesar de as características destes modelos se ajustarem bastante bem ao comportamento dos custos com sinistros de diversos seguros, como é o caso da Responsabilidade Civil Automóvel, existem outros casos em que esse ajustamento já não é tão bom.

Neste trabalho estudam-se duas características de alguns seguros, que limitam a adequação dos modelos lineares generalizados para modelizar os custos com sinistros: as franquias e os limites de indemnização. A existência de franquias implica que apenas sejam participados sinistros cujo custo não exceda o valor da franquia, enquanto que a aplicação de limites de indemnização conduz a que não se conheça o valor do dano, quando este excede o limite de indemnização, porque apenas fica registado o valor indemnizado.

Para superar a inadequação dos modelos lineares generalizados na modelização destas situações, propõe-se a utilização dos modelos tobit generalizados. Estes modelos, que seguem a filosofia dos modelos tobit, permitem estimar a distribuição dos custos subjacentes às indemnizações, corrigindo o efeito da existência de franquias e de limites de indemnização.

Com o intuito de ilustrar a utilização destes modelos, apresenta-se uma aplicação à cobertura de Choque, Colisão ou Capotamento, do seguro automóvel.

Como as franquias têm impacto no número de sinistros participados, é apresentado um procedimento para estimar o número de sinistros ocorridos, independentemente do valor das franquias.

Palavras-chave: modelos lineares generalizados; modelos tobit; franquias; limites de indemnização; tarifação.

Abstract

The construction of premium rating structures is, usually, made in two steps: average claim amount estimation and frequency claim estimation. Generalised linear models are used in a generic way when building those structures, for kinds of insurance with a considerable number of risk units and a low level of heterogeneity among them, being able to estimate any of those parts.

Although the characteristics of these models fit well to the behaviour of claim costs of several insurance kinds, such as the third party motor insurance, there are other cases where it doesn't fit so well.

In this thesis are studied two characteristics of some insurance kinds that bounds the adequacy of generalised linear models: deductibles and policy limits. Because of the existence of deductibles, only are notified those claims whose monetary loss doesn't exceed the deductible amount, while the policy limit implies that the monetary loss is unknown when it exceeds that limit.

To overcome the inadequacy of generalised linear models under these constraints, it is proposed the use of generalised tobit models. These models, that follow the philosophy of tobit models, allow to estimate the distribution of monetary losses that originates the claim amount, correcting the effects of deductibles and policy limits.

With the aim of illustrating the use of these models, it is shown an application to the own damage collision cover of motor insurance.

As the deductibles also have impact on the number of notified claims, it is presented a procedure to estimate the number of occurred claims, independently of the deductible amount.

Keywords: generalised linear models; tobit models; deductibles; policy limits; rating structures.

Agradecimentos

Os meus agradecimentos dirigem-se a todos aqueles que, de várias formas, me ajudaram e contribuíram para este trabalho.

Gostaria de começar por agradecer ao Professor Doutor João Andrade e Silva, que orientou o desenvolvimento desta dissertação. A sua disponibilidade foi total e os seus contributos foram de inquestionável importância para melhorar a qualidade deste trabalho.

Quero agradecer ao Sr. Luís Caldas por me ter motivado e por ter criado as condições profissionais para que eu pudesse frequentar o mestrado, para além de toda a experiência e conhecimentos que me transmitiu.

Às minhas colegas dos gabinetes de Actuariado da Sociedade Portuguesa de Seguros e da Allianz-Portugal, a Dra. Isabel Ribeiro e a Dra. Concórdia Simão, que sempre me encorajaram e se disponibilizaram para discutir aspectos deste estudo.

Quero também agradecer ao Eng. Carlos Coutinho, à Dra. Luisa Santos, ao Dr. Manuel Taveira e aos restantes membros da Direcção de Produtos, da Mundial-Confiança, nomeadamente às colegas do Dep. de Estudos e Actuariado, pelo apoio e esclarecimentos prestados.

À minha família e amigos pela paciência que demonstraram para comigo, apoiando-me sempre e incondicionalmente.

À Graça pelo apoio, companhia, ajuda e compreensão que demonstrou ao longo de todo este tempo.

Muito obrigado.

Conteúdo

1 Introdução 9

2 A abordagem habitual - os modelos lineares generalizados 13

2.1 Introdução 13

2.2 As componentes dos modelos lineares generalizados 13

2.3 Estimação dos parâmetros 15

2.3.1 Funções de verosimilhança para modelos lineares generalizados . . . 15

2.3.2 Estimação do vector β 17

2.3.3 Estimação do parâmetro de escala 19

2.4 A qualidade do ajustamento 20

3 Uma abordagem alternativa - os modelos tobit generalizados 24

3.1 Introdução 24

3.2 Modelos de regressão com variáveis dependentes limitadas 24

3.2.1 Variáveis censuradas e truncadas 24

3.2.2 O modelo de regressão tobit 26

3.3 A generalização dos modelos tobit 31

3.3.1 Funções de verosimilhança e estimação 31

3.3.2 Inferência sobre os parâmetros 34

3.4 Alguns exemplos de modelos tobit generalizados 35

3.4.1 A modelização com distribuição lognormal 35

3.4.2 A modelização com distribuição gama 41

3.4.3 A modelização com distribuição inversa Gaussiana 50

4 Limitações e pontencialidades das duas abordagens 58

4.1 Franquias e limites de indemnização 60

4.1.1 Franquias 60

4.1.2 Limites de indemnização 64

4.2 Insuficiências dos modelos lineares generalizados 66

4.3 A modelização dos custos com os modelos tobit generalizados 68

5 Estimação da distribuição de custos com sinistros **72**

5.1 Características da cobertura de Choque, Colisão ou Capotamento 72

5.2 A informação utilizada 75

5.2.1 Características dos sinistros 75

5.2.2 Características da base de dados 78

5.2.3 Variáveis analisadas 79

5.3 Resultados de uma modelização genérica 84

5.4 Resultados de uma modelização alternativa 93

5.4.1 Pressupostos da abordagem alternativa 94

5.4.2 A modelização das perdas parciais 95

5.4.3 A modelização das perdas totais 98

5.4.4 A modelização da probabilidade de perda total 100

5.5 Comparação dos resultados das duas modelizações 103

5.5.1 Resultados para alguns exemplos 103

5.5.2 Avaliação da qualidade dos ajustamentos 106

6 A interacção entre frequência e custo **109**

6.1 A estimação da frequência 109

6.2 A combinação da frequência e do custo 112

7 Conclusões e comentários finais **114**

8 Bibliografia **118**

9 Anexos 120

Lista de Figuras

1 Prémios brutos emitidos de seguro directo em 2000 (valores em milhares de escudos) 11

2 Resultados de simulação com distribuição lognormal 40

3 N°de observações censuradas e truncadas - distribuição lognormal 41

4 Erros das aproximações da função de distribuição gama 47

5 Resultados de simulação com distribuição gama 50

6 N°de observações censuradas e truncadas - distribuição gama 51

7 Resultados de simulação com distribuição inversa Gaussiana 56

8 N°de observações censuradas e truncadas - distribuição inversa Gaussiana . 57

9 Estimativas dos parâmetros de modelos truncados e censurados (valores em escudos) 88

10 Qualidade do ajustamento de modelos truncados e censurados com diversas distribuições 92

11 Número estimado de sinistros com custo superior a 70% do capital seguro . 93

12 Estimativas do modelo de custos de perdas parciais (valores em escudos) . . 96

13 Estimativas do modelo de probabilidade de perda total 102

14 Valores esperados das indemnizações resultantes das duas modelizações (valores em escudos) 105

15 Probabilidades de custo com sinistro inferior à franquia 105

16 Probabilidades de perda total 106

17 Número estimado de sinistros por intervalo de custo 108

1 Introdução

A indústria seguradora tem assistido, nos últimos anos, a um aumento significativo da concorrência. As seguradoras a operar no mercado português têm direccionado uma parte importante dos seus esforços no sentido de aumentar as suas carteiras de apólices e, consequentemente, o montante de prémios cobrados.

Para alcançar esse objectivo, as seguradoras têm utilizado diversas estratégias, que passam muitas vezes pela prestação de melhores serviços aos segurados e por um esforço de criar uma boa imagem junto de potenciais clientes. No entanto, por ser, em muitos casos, difícil criar produtos diferenciados, o factor mais determinante para o sucesso da comercialização de um seguro é o seu preço. Assim, o acentuar da concorrência tem conduzido a uma tendência para a redução dos prémios, com a consequente quebra de resultados.

Nestas condições, é de extrema importância o conhecimento do comportamento da sinistralidade dos diversos segmentos de risco. Só assim se podem construir tarifas que tornem o produto atractivo para os segmentos mais rentáveis e que permitam alcançar os objectivos das seguradoras: o aumento da quota de mercado sem pôr em causa a rentabilidade da empresa.

Os procedimentos de análise do comportamento da sinistralidade não são uniformes para todos os tipos de seguros. As ferramentas mais adequadas para efectuar esse estudo dependem das características do seguro, da forma como são determinadas as indemnizações, da maior ou menor heterogeneidade entre os diversos riscos existentes em carteira, etc.. Por exemplo, as metodologias estatísticas utilizadas na análise tarifária de seguros de acidentes de trabalho e de responsabilidade civil automóvel são substancialmente diferentes. No caso de seguros relacionados com riscos industriais, é inviável a análise da sinistralidade com o objectivo de construir uma tarifa, não só porque a experiência da

seguradora é insuficiente, mas também porque cada risco tem características muito específicas, sendo habitual definir o prémio de forma casuística, com base na apreciação de um analista de risco.

As metodologias de tarifação a priori são principalmente aplicadas em seguros que se caracterizam por serem direccionados para clientes particulares, com um número considerável de apólices e com um reduzido grau de heterogeneidade entre os riscos.

O seguro de responsabilidade civil automóvel é aquele em que as seguradoras sentem uma maior necessidade de construir uma tarifa segmentada e ajustada aos comportamentos da sinistralidade. Essa preocupação resulta de os prémios desse seguro representarem cerca de 2/3 do total de prémios do ramo Automóvel, do mercado português no ano 2000, ou seja, aproximadamente 32% do total de prémios dos ramos Não Vida (Figura 1).

Apesar de ser esse o produto de maior peso, diversos outros seguros, que contribuem significativamente para a composição das carteiras de contratos, permitem também a estimação de estruturas tarifárias. De entre esses seguros destacam-se: as coberturas de danos próprios do seguro automóvel, que representam cerca de 16% do total de prémios (1/3 dos prémios do ramo Automóvel), os seguros de Doença (6,14% dos prémios), os seguros de Riscos Múltiplos Habitação (5,78% dos prémios), outros seguros do ramo de Incêndio e Outros Danos e alguns seguros de Responsabilidade Civil Geral.

Quando se procede à modelização de estruturas tarifárias para esses seguros, os actuários recorrem frequentemente aos modelos lineares generalizados. No entanto, vários seguros referenciados possuem características, relacionadas com franquias e limites de indemnização, que tornam algo desajustada a utilização dos modelos lineares generalizados, sendo por isso necessário procurar um método mais adequado para esses casos.

Com o objectivo de ultrapassar as limitações dos modelos lineares generalizados, propõe-se a utilização de modelos alternativos de estimação de estruturas tarifárias, baseados nos

Ramos Não Vida				Prémios	Proporção
Acidentes e Doença				197 687 774	30,12%
	Acidentes			157 366 185	23,97%
		Acidentes de Trabalho		128 576 433	19,59%
		Acidentes Pessoais		24 194 731	3,69%
		Pessoas Transportadas		4 595 021	0,70%
	Doença			40 321 589	6,14%
Incêndio e Outros Danos				99 272 145	15,12%
	Incêndio e Elem. da Natureza			11 277 367	1,72%
	Outros Danos em Coisas			87 994 778	13,41%
		Agrícola		9 566 673	1,46%
		Roubo		2 006 645	0,31%
		Avaria de Máquinas		3 133 788	0,48%
		Riscos Múltiplos		67 623 942	10,30%
			Habitação	37 941 024	5,78%
			Comerciantes	18 335 705	2,79%
			Industrial	7 729 006	1,18%
		Outros		5 663 730	0,86%
Automóvel				319 622 422	48,69%
Marítimo e Transportes				3 895 729	0,59%
Aéreo				1 760 824	0,27%
Mercadorias Transportadas				6 871 159	1,05%
Responsabilidade Civil Geral				11 350 360	1,73%
Diversos				15 948 408	2,43%

Figura 1: Prémios brutos emitidos de seguro directo em 2000 (valores em milhares de escudos)

modelos tobit. Com esses modelos, designados por modelos tobit generalizados, é possível estimar a distribuição dos custos que dão origem às indemnizações, superando os problemas de modelização associados às franquias e aos limites de indemnização. Estes modelos permitem estimar o impacto de diversos tipos de franquias e limites de indemnização sobre o valor esperado das indemnizações a cargo da seguradora.

Para ilustrar a forma como podem ser utilizados, bem como os resultados que fornecem, aplicam-se os modelos tobit generalizados à cobertura de Choque, Colisão ou Capotamento, a cobertura com maior peso no seguro de danos próprios automóvel. A sua contratação pressupõe sempre a definição de um limite de indemnização e, quase sempre, a aplicação de uma franquia.

Este trabalho está estruturado da seguinte forma: no capítulo seguinte apresentam-se sucintamente os modelos lineares generalizados; no Capítulo 3 desenvolve-se a metodologia

alternativa, recorrendo aos modelos tobit generalizados; no Capítulo 4 focam-se as características dos seguros que condicionam a adequação de cada uma das metodologias; após a apresentação das metodologias, aplicam-se os modelos propostos à cobertura de Choque, Colisão ou Capotamento do seguro automóvel; no Capítulo 6 expõem-se os impactos das franquias sobre o comportamento da frequência de sinistralidade; por fim, no 7°capítulo referem-se as principais conclusões do trabalho e tecem-se algumas considerações finais.

2 A abordagem habitual - os modelos lineares generalizados

2.1 Introdução

A metodologia habitualmente utilizada na estimação de tarifas baseia-se nos modelos lineares generalizados. Estes modelos permitem estimar o impacto de diversas variáveis, os factores de tarifação, sobre o valor esperado dos dois principais aspectos da sinistralidade das apólices de seguro: a frequência de sinistralidade e a severidade de cada sinistro. As principais vantagens destes modelos relativamente a outros, nomeadamente ao modelo de regressão linear, são a possibilidade de alargar as famílias de distribuições disponíveis e o facto de o valor esperado condicionado não ser necessariamente uma função linear das variáveis explicativas. Nesta secção serão apresentadas as principais características dos modelos lineares generalizados. Uma apresentação mais detalhada destes modelos pode ser encontrada em McCullagh e Nelder (1989).

2.2 As componentes dos modelos lineares generalizados

Os modelos lineares generalizados podem ser entendidos como uma extensão do modelo de regressão linear clássico. Assume-se que as observações y_i $(i = 1, ..., n)$ são realizações, com distribuição independente, de variáveis aleatórias Y_i com valor esperado condicionado μ_i. No modelo de regressão linear considera-se que o valor esperado de cada observação é condicionado por p variáveis explicativas x_{ij} $(j = 1, ..., p)$ e que essa relação assume a forma

$$E(Y_i|x_{i1}, ..., x_{ip}) = \mu_i = \sum_{j=1}^{p} x_{ij}\beta_j \qquad i = 1, ..., n, \tag{1}$$

onde os β_j são parâmetros cujo valor é normalmente desconhecido. Com o objectivo de simplificar a notação, daqui em diante utilizar-se-á $E(Y_i)$ para indicar o valor esperado

condicionado. A expressão (1) especifica apenas o comportamento de $E(Y_i)$, sendo também necessário definir as restantes características da distribuição de Y_i. No modelo linear classico assume-se que Y_i tem distribuição normal com variância constante σ^2.

Os modelos lineares generalizados, introduzidos por Nelder e Wedderburn (1972), constituem uma generalização do modelo de regressão linear clássico, apresentando hipóteses menos restritivas em dois aspectos. Primeiro, a distribuição da variável aleatória não é forçosamente normal, podendo ser uma outra distribuição da família exponencial. Em segundo lugar, o valor esperado não é necessariamente uma função linear das variáveis explicativas.

Um modelo linear generalizado é caracterizado fundamentalmente por:

1. A distribuição de Y_i pertence a uma família de distribuições, incluída na família de dispersão exponencial, caracterizada pelos parâmetros θ_i e ϕ, e tal que a sua função densidade ou a função de probabilidade é

$$f(y_i;\theta_i,\phi) = \exp\left[\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i,\phi)\right],$$

onde θ é o parâmetro de localização e está indexado pela observação i, e ϕ é o parâmetro de escala que é fixo para todas as observações. As funções $b(\cdot)$ e $c(\cdot,\cdot)$ não variam com a observação, ao contrário de $a_i(\cdot)$, que se assume como sendo $a_i(\phi) = \phi/\omega_i$, onde ω_i é um ponderador conhecido *a priori* que varia de observação para observação.

2. A existência de uma estrutura linear, usualmente designada de preditor linear, resultante da combinação de p variáveis explicativas, dada por

$$\eta_i = \sum_{j=1}^{p} x_{ij}\beta_j \qquad i = 1, ..., n,$$

onde x_{ij} é a i-ésima observação da variável explicativa j $(j = 1, ..., p)$. Por forma a que o modelo tenha termo independente, considera-se que $x_{i1} = 1$, para todas as i observações.

3. A existência de uma função $g(\cdot)$, monótona e diferenciável, designada como função de ligação, que estabelece a relação entre μ_i, o valor esperado da variável Y_i, e o preditor linear η_i

$$\eta_i = g(\mu_i) \qquad i = 1, ..., n.$$

2.3 Estimação dos parâmetros

2.3.1 Funções de verosimilhança para modelos lineares generalizados

Assume-se que cada observação de Y segue uma distribuição da família exponencial, tendo a forma de

$$f(y; \theta, \phi) = \exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right], \tag{2}$$

para algumas funções específicas de $b(\cdot)$, $a(\cdot)$ e $c(\cdot, \cdot)$.

Seja $l(\theta, \phi; y) = \ln f(y; \theta, \phi)$ a função de log-verosimilhança considerada como função de θ e ϕ, dado y. A média e a variância de Y podem ser deduzidas através das bem conhecidas relações,

$$E\left[\frac{\partial l}{\partial \theta}\right] = 0 \tag{3}$$

e

$$E\left[\frac{\partial l}{\partial \theta}\right]^2 + E\left[\frac{\partial^2 l}{\partial \theta^2}\right] = 0 \tag{4}$$

cuja demonstração pode ser encontrada, por exemplo, em Dobson (1990).

De (2) temos que

$$l(\theta, \phi; y) = \frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\,. \tag{5}$$

Considerando a expressão (5) e as relações (3) e (4) tem-se que, como se mostra em McCullagh e Nelder (1989),

$$E[Y] = \mu = b'(\theta)\,, \qquad Var[Y] = b''(\theta)\, a(\phi)\,.$$

Verifica-se assim que o valor esperado μ depende apenas do parâmetro θ, enquanto a variância de Y é o produto de duas parcelas: a primeira, $b''(\theta)$, depende de θ, será designada de função variância e escreve-se $V(\mu)$; a segunda depende apenas de ϕ.

As distribuições mais importantes da forma (2) encontram-se sumarizadas no quadro seguinte,

	Normal	Poisson	Binomial	Gama	Inversa Gaussiana
Notação	$N(\mu, \sigma^2)$	$P(\mu)$	$B(m, \mu)/m$	$G(\mu, \nu)$	$IG(\mu, \sigma^2)$
ϕ	σ^2	1	$1/m$	ν^{-1}	σ^2
$b(\theta)$	$\theta^2/2$	$\exp(\theta)$	$\ln(1+e^{\theta})$	$-\ln(-\theta)$	$-(-2\theta)^{1/2}$
$c(y, \phi)$	$-\frac{1}{2}\frac{y^2}{\sigma^2} +$ $-\frac{1}{2}\ln(2\pi\sigma^2)$	$-\ln y!$	$\ln\begin{pmatrix} m \\ my \end{pmatrix}$	$\nu\ln(\nu y) +$ $-\ln y - \ln\Gamma(\nu)$	$-\frac{1}{2}\ln(2\pi\sigma^2 y^3)$ $-\frac{1}{2}\frac{1}{\sigma^2 y}$
$\mu(\theta)$	θ	$\exp(\theta)$	$\frac{e^{\theta}}{1+e^{\theta}}$	$-1/\theta$	$(-2\theta)^{-1/2}$
link canónico	μ	$\ln\mu$	$\ln\left(\frac{\mu}{1-\mu}\right)$	$1/\mu$	$1/\mu^2$
$V(\mu)$	1	μ	$\mu(1-\mu)$	μ^2	μ^3

Entende-se por link canónico o link tal que $\eta = g(\mu) = \theta$.

2.3.2 Estimação do vector $\boldsymbol{\beta}$

A estimação dos parâmetros β_j $(j = 1, ..., p)$ é efectuada recorrendo à maximização da função de log-verosimilhança, dada a amostra observada,

$$l = \sum_{i=1}^{n} l_i(\beta_1, ..., \beta_p, \phi; y_i, x_{i1}, ..., x_{ip}) = \sum_{i=1}^{n} \left[\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi)\right],$$

já que θ_i é função dos parâmetros β_j, uma vez que $g(b'(\theta_i)) = g(\mu_i) = \sum_{j=1}^{p} x_{ij}\beta_j$.

Utilizando o método dos scores de Fisher para a estimação do vector de parâmetros $\boldsymbol{\beta}$, um método iterativo da família de algoritmos de Newton-Raphson, tem-se para a m-ésima iteração

$$\boldsymbol{\beta}^{(m)} = \boldsymbol{\beta}^{(m-1)} + E\left[-\frac{\partial^2 l}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'}\right]^{-1}_{\boldsymbol{\beta}^{(m-1)}} \left[\frac{\partial l}{\partial\boldsymbol{\beta}}\right]_{\boldsymbol{\beta}^{(m-1)}},$$

o que obriga a que se calcule $\left[\frac{\partial l}{\partial\boldsymbol{\beta}}\right]$ e $E\left[-\frac{\partial^2 l}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'}\right]$.

O vector $\left[\frac{\partial l}{\partial \beta}\right]$ tem elemento genérico

$$\frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial l_i}{\partial \beta_j}.$$

Dadas as relações entre as diversas funções, tem-se

$$\sum_{i=1}^{n} \frac{\partial l_i}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial l_i}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j}.$$

Calculando as diversas derivadas, verifica-se que

$$\frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{n} x_{ij} \tilde{z}_i,$$

onde

$$\tilde{z}_i = \frac{Y_i - \mu_i}{Var\,(Y_i)} \frac{\partial \mu_i}{\partial \eta_i}.$$

No que se refere a $E\left[-\frac{\partial^2 l}{\partial \beta \partial \beta}\right]$, o seu elemento genérico é dado por

$$E\left[-\frac{\partial^2 l}{\partial \beta_j \partial \beta_k}\right] = \sum_{i=1}^{n} x_{ij} W_{ii} x_{ik},$$

com

$$W_{ii} = \frac{1}{Var\,(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2.$$

Se definirmos $\mathbf{W}$ como a matriz diagonal com elemento genérico W_{ii}, vem

$$E\left[-\frac{\partial^2 l}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'}\right] = \mathbf{X}'\mathbf{W}\mathbf{X}.$$

Assim, a estimação de $\boldsymbol{\beta}$ é feita pelo seguinte método iterativo:

$$\begin{aligned}\hat{\boldsymbol{\beta}}^{(m)} &= \hat{\boldsymbol{\beta}}^{(m-1)} + \left(\mathbf{X}'\mathbf{W}^{(m-1)}\mathbf{X}\right)^{-1}\mathbf{X}'\tilde{\mathbf{z}}^{(m-1)} \\ &= \left(\mathbf{X}'\mathbf{W}^{(m-1)}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{W}^{(m-1)}\mathbf{z}^{(m-1)}, \end{aligned} \tag{6}$$

com $z_i = \eta_i + \frac{1}{W_{ii}}\tilde{z}_i = \eta_i + (y_i - \mu_i)\frac{\partial \eta_i}{\partial \mu_i}$.

Regra geral, em cada iteração, a matriz $\mathbf{W}$ e o vector $\mathbf{z}$ dependem da estimativa $\hat{\boldsymbol{\beta}}$.

A expressão mostra que as estimativas de máxima verosimilhança dos parâmetros $\boldsymbol{\beta}$ são obtidas pelo método iterativo dos mínimos quadrados ponderados. Mais detalhes sobre esta metodologia podem ser encontrados em McCullagh e Nelder (1989).

2.3.3 Estimação do parâmetro de escala

Após ter estimado o vector $\boldsymbol{\beta}$, falta ainda estimar o parâmetro de escala ϕ. A sua estimação pelo método da máxima verosimilhança, embora possível, é complicada devido à inexistência de uma solução geral explícita para as distribuições da família exponencial.

Dada a impossibilidade de uma metodologia uniforme para gerar as estimativas de máxima verosimilhança para ϕ, recorre-se a um estimador baseado na *deviance* à escala (D^*), uma estatística que será apresentada na próxima secção, ou na estatística X^2 de Pearson.

Como se sabe que em termos assimptóticos e num quadro de hipóteses bastante geral $E[D^*] = n - p$, então um estimador assimptoticamente centrado para ϕ é

$$\hat{\phi} = \frac{D}{n-p}. \tag{7}$$

Quando se usa o X^2 de Pearson a expressão é semelhante

$$\hat{\phi} = \frac{X^2}{n-p}, \tag{8}$$

onde X^2 é definido como

$$X^2 = \sum_i \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}, \tag{9}$$

onde $V(\hat{\mu}_i)$ é o valor da função variância adequada para a distribuição em uso, no ponto $\mu_i = \hat{\mu}_i$.

2.4 A qualidade do ajustamento

As medidas de qualidade de ajustamento servem para medir a discrepância entre um conjunto de observações y_i e um conjunto de valores ajustados $\hat{\mu}_i$. Estas medidas podem ser de diversos tipos, mas em termos dos modelos lineares generalizados é geralmente utilizada uma medida construída a partir do logaritmo de uma razão de verosimilhanças, designada *deviance*.

Dadas n observações, podem-se ajustar modelos que tenham até n parâmetros. O modelo mais simples, o *modelo nulo* ou *mínimo*, tem apenas um parâmetro, o termo independente, e pressupõe que todas as observações têm um valor esperado comum, o que implica que se assume toda a variação entre observações como sendo aleatória. O *modelo completo* ou *saturado* situa-se no extremo oposto: tem n parâmetros, o que implica que o valor estimado de cada observação coincide com o valor realmente observado. Assim, no modelo completo toda a variação entre as observações é considerada como sistemática.

Dadas as suas características, os modelos nulo e completo são, na prática, pouco utilizados. O primeiro é demasiado simples, tendo pouco poder explicativo do comportamento

de fenómenos com alguma complexidade, enquanto o segundo, por ter um elevado número de parâmetros, não permite sintetizar o impacto das principais variáveis exógenas. Apesar do reduzido interesse destes modelos, com um número extremo de parâmetros, para explicar um fenómeno, eles são bastante úteis para auxiliar a medição da discrepância do ajustamento.

O modelo completo, ao implicar o ajustamento total do modelo aos dados, apresenta o maior valor que a função de verosimilhança pode assumir para determinado conjunto de observações, fornecendo um valor de referência para medir a discrepância de um modelo parcimonioso.

Define-se então *deviance à escala* como sendo duas vezes a diferença entre a máxima log-verosimilhança alcançável e aquela que é alcançada pelo modelo em estudo. Assim, a *deviance à escala*, para um valor fixo de ϕ, pode ser escrita como

$$\begin{aligned} D^*\left(y;\tilde{\theta}\right) &= 2\left[l\left(\tilde{\theta},\phi;y\right) - l\left(\hat{\theta},\phi;y\right)\right] \\ &= \sum_i \frac{2\omega_i}{\phi}\left\{y_i\left(\tilde{\theta}_i - \hat{\theta}_i\right) - \gamma\left(\tilde{\theta}_i\right) + \gamma\left(\hat{\theta}_i\right)\right\} \\ &= \frac{D\left(y;\tilde{\theta}\right)}{\phi}, \end{aligned}$$

onde $\hat{\theta}$ e $\tilde{\theta}$ são as estimativas do modelo em estudo e do modelo saturado, respectivamente. $D\left(y;\tilde{\theta}\right)$ é conhecida como *deviance* e depende apenas das observações, ao contrário do que sucede com a *deviance à escala*, que também depende de ϕ.

Uma medida alternativa de discrepância é dada pela estatística X^2 de Pearson generalizada, tal como foi definida em (9).

A *deviance à escala*, num modelo adequadamente especificado tem uma distribuição assimptótica que, em condições relativamente gerais, se aproxima de uma qui-quadrado não central com $n-p$ graus de liberdade, como se pode ver em McCullagh e Nelder (1989).

Quando se trata de um modelo com distribuição normal e função de ligação identidade, a distribuição é exacta mas, nos restantes casos, as inferências baseadas nesta estatística são muito limitadas.

Se considerarmos dois modelos com a mesma variável endógena, mesma distribuição e mesma função de ligação, em que os p_2 parâmetros do segundo são o resultado de $p_1 - p_2$ restrições lineares sobre os p_1 parâmetros do primeiro modelo, então diz-se que os modelos são encaixados. Sendo D_1^* e D_2^* as deviance à escala do primeiro e do segundo modelo, respectivamente, tem-se que $D_1^* \leqslant D_2^*$. Facilmente se mostra, assumindo que o modelo 2 é o correcto, que

$$\Delta D^* = D_2^* - D_1^* = \frac{D_2 - D_1}{\phi} \dot{\sim} \chi^2_{(p_1 - p_2)}. \tag{10}$$

Note-se que, ao contrário do que sucede com a deviance à escala, ΔD^* tem distribuição do qui-quadrado e não uma qui-quadrado não central. Para além disso, a convergência de ΔD^* para a sua distribuição assimptótica é rápida. Este resultado é de grande utilidade, permitindo efectuar testes de hipóteses sobre restrições lineares, nomeadamente de nulidade, a conjuntos de parâmetros do primeiro modelo. O procedimento habitual para efectuar os testes de hipóteses consiste em estimar determinado modelo, estimar esse modelo reespecificado por forma a incorporar as $(p_1 - p_2)$ restrições lineares, e calcular a diferença entre as deviance dos dois modelos.

A utilização de (10) pressupõe que o parâmetro ϕ é conhecido. Quando não se conhece ϕ, pode-se utilizar o resultado anterior, substituindo ϕ por uma sua estimativa consistente obtida, nomeadamente, através de (7) ou (8), ou então recorrer ao seguinte resultado

(Dobson (1990))

$$\frac{\frac{D_2-D_1}{p_1-p_2}}{\frac{D_1}{n-p_1}} \dot{\sim} F_{(p_1-p_2;n-p_1)},$$

obtido assumindo também que o modelo 2 é o correcto.

A avaliação da qualidade global do modelo pode ser realizada utilizando os resultados anteriores para comparar o modelo em análise com o modelo nulo, que pode sempre ser entendido como resultando da imposição de $(p-1)$ restrições lineares sobre o modelo em estudo. Assim, poderá testar-se a significância global dos parâmetros.

A inferência sobre um parâmetro individual poderá ser efectuada através da comparação de modelos encaixados, com apenas uma restrição no modelo com menos parâmetros, ou alternativamente, recorrendo à distribuição assimptótica do estimador de máxima verosimilhança, que se sabe ser

$$\hat{\beta}_j \dot{\sim} N\left(\beta_j, \nu_{jj}\right),$$

onde ν_{jj} é o elemento (j,j) da inversa da matriz de informação de Fisher.

3 Uma abordagem alternativa - os modelos tobit generalizados

3.1 Introdução

No Capítulo 2 expôs-se a metodologia que habitualmente se utiliza para analisar estatisticamente o processo de custos com sinistros de uma grande diversidade de seguros. Neste capítulo é apresentada uma proposta alternativa para a modelização dos custos dos sinistros. A necessidade de se propor uma metodologia alternativa aos modelos lineares generalizados, resulta de estes não permitirem modelizar correctamente alguns aspectos das indemnizações de determinados tipos de seguros. Essas insuficiências, dos modelos lineares generalizados, serão abordadas do Capítulo 4 em diante.

A metodologia proposta para modelizar os custos dos sinistros deriva em grande parte da filosofia dos modelos lineares generalizados, cruzando algumas das suas características com os modelos tobit, os quais foram inicialmente estudados por James Tobin (1958). Na secção seguinte indicam-se os objectivos e alguns resultados dos modelos tobit. A sua apresentação, terminologia e notação seguem de perto Maddala (1983).

3.2 Modelos de regressão com variáveis dependentes limitadas

Os modelos tobit foram desenvolvidos com o objectivo de construir modelos de regressão para os quais a variável dependente, com distribuição normal, apenas é observada dentro de determinado intervalo. A não observação de alguns valores pode resultar de dois fenómenos: a censura e a truncagem.

3.2.1 Variáveis censuradas e truncadas

Suponha-se que Y^* é uma variável aleatória com distribuição normal, de média μ e variância σ^2 e considere-se uma amostra de dimensão n, $(y_1^*, y_2^*, ..., y_n^*)$, da qual só são registados

os valores de y^* que são superiores a uma constante c. Para valores de $y^* \leqslant c$, regista-se o valor c. Assim, os valores observados na amostra são

$$y_i = \begin{cases} y_i^* & \text{se } y_i^* > c \\ c & \text{caso contrário} \end{cases} .$$

A amostra obtida, $(y_1, y_2, ..., y_n)$, denomina-se de *amostra censurada à esquerda*. Para as observações $y_i = c$, tudo o que se sabe é que $y_i^* \leqslant c$, ou seja,

$$\Pr(Y_i = c) = \Pr(Y_i^* \leqslant c) .$$

Assim, a função de verosimilhança para a estimação dos parâmetros μ e σ^2, observada uma amostra de dimensão n, vem

$$L(\mu, \sigma^2 | y) = \prod_{y_i^* > c} \sigma^{-1} \phi\left(\frac{y_i - \mu}{\sigma}\right) \cdot \prod_{y_i^* \leqslant c} \Phi\left(\frac{c - \mu}{\sigma}\right),$$

onde $\phi(\cdot)$ e $\Phi(\cdot)$ são, respectivamente, as funções densidade e de distribuição da normal estandardizada.

Suponhamos agora que, na recolha da amostra, não são observáveis valores de $y^* < c$. Diz-se então que as observações da amostra resultam de uma *variável truncada* à esquerda em c. A função densidade da distribuição normal truncada é dada pela seguinte expressão:

$$f(y^* | Y^* \geqslant c) = \frac{\sigma^{-1}\phi\left(\frac{y^* - \mu}{\sigma}\right)}{1 - \Phi\left(\frac{c - \mu}{\sigma}\right)} \qquad c \leqslant y^* < +\infty.$$

O denominador da expressão é a constante de normalização, pois o integral do numerador em todo o domínio da variável é igual a $1 - \Phi\left(\frac{c-\mu}{\sigma}\right)$.

Uma amostra recolhida de uma distribuição com estas características é denominada

de amostra truncada.

Podem-se considerar amostras com diversas combinações de truncagens e/ou censuras. Por exemplo, podemos ter uma amostra truncada-censurada, considerando a truncagem num valor c_1 e a censura num valor c_2, com $c_1 < c_2$, por exemplo. Assim, só são observados valores de $y^* \geqslant c_1$ e, destas observações, apenas são registadas aquelas em que $y^* < c_2$; quando $y^* \geqslant c_2$, o valor registado é c_2; ou seja,

$$y_i = \begin{cases} y_i^* & \text{se } c_1 \leqslant y_i^* < c_2 \\ c_2 & \text{se } y_i^* \geqslant c_2 \end{cases}.$$

A função de verosimilhança para um modelo com estas características é

$$L\left(\mu, \sigma^2 | y\right) = \left[1 - \Phi\left(\frac{c_1 - \mu}{c}\right)\right]^{-n} \prod_{y_i^* < c_2} \sigma^{-1} \phi\left(\frac{y_i - \mu}{\sigma}\right) \prod_{y_i^* \geqslant c_2} \Phi\left(\frac{c_2 - \mu}{\sigma}\right).$$

3.2.2 O modelo de regressão tobit

O modelo de regressão tobit propõe-se estimar o comportamento de variáveis aleatórias para as quais alguns valores não são observáveis no processo de recolha da amostra. A impossibilidade de se observarem determinados valores da variável aleatória pode ter origem em fenómenos de censura e/ou truncagem. Considere-se que Y_i^* é a variável aleatória que se pretende modelizar e que

$$Y_i^* \sim N\left(\sum_{j=1}^{k} x_{ij}\beta_j, \sigma^2\right) \qquad i = 1, ..., n,$$

onde x_{ij} $(j = 1, ..., k)$ é a i-ésima observação da variável explicativa j, e os β_j são parâmetros desconhecidos.

Considere-se também que, devido à existência de censura à esquerda no valor c, alguns valores de Y_i^* não são observáveis. Como resultado da censura, os valores de Y_i^* que se

observam são

$$y_i = \begin{cases} y_i^* & \text{se } y_i^* > c \\ c & \text{caso contrário} \end{cases}.$$

Assumindo que c é conhecido, pode-se sempre reescrever o modelo por forma a obter um modelo equivalente com ponto de censura em zero. Essa transformação é feita subtraindo c à variável endógena e ao termo independente do modelo. Dada a possibilidade de passar de um modelo com ponto de censura qualquer, para um modelo com censura no ponto zero, apenas se tratará, daqui em diante, o caso em que $c = 0$.

Nestas condições, o modelo fica

$$E(Y_i) = \begin{cases} E(Y_i^* | Y_i^* > 0) & \text{se } Y_i^* > 0 \\ 0 & \text{caso contrário} \end{cases}.$$

Refira-se que os modelos com ponto de censura à direita são tratados de forma semelhante aos modelos com censura à esquerda.

Uma vez que a estimação da distribuição de Y_i^*, utilizando o método dos mínimos quadrados com base numa amostra censurada y_i, originaria estimativas enviesadas, um dos objectivos dos modelos tobit é a estimação centrada dos parâmetros β_j e σ^2.

Um aspecto importante destes modelos é que devem ser sempre conhecidos os valores das variáveis explicativas x_{ij}, mesmo quando se observa $y_i = 0$, isto é, quando as observações da variável endógena são censuradas.

Assuma-se que n_0 é o número de observações para as quais $y_i = 0$, que n_1 é o número de observações em que $y_i > 0$ e que as observações são ordenadas por forma a que as n_0 primeiras observações correspondam aos valores de $y_i = 0$.

Por conveniência, defina-se F_i e f_i como sendo as funções de distribuição e de densidade

de probabilidade da normal estandardizada, respectivamente, avaliadas no ponto $\sigma^{-1}\mu_i$, com $\mu_i = \sum_{j=1}^{k} x_{ij}\beta_j$, isto é, $F_i = \Phi\left(\frac{\mu_i}{\sigma}\right)$ e $f_i = \phi\left(\frac{\mu_i}{\sigma}\right)$. Para as observações $y_i = 0$ tudo o que se sabe é que $y_i^* \leqslant 0$, então

$$\Pr(Y_i = 0) = \Pr(Y_i^* \leqslant 0) = \Pr\left(\frac{y_i^* - \mu_i}{\sigma} \leqslant \frac{-\mu_i}{\sigma}\right) = (1 - F_i),$$

uma vez que a distribuição normal é simétrica. Para as observações em que $y_i > 0$, tem-se que

$$\Pr(Y_i > 0) \cdot f(y_i | Y_i > 0) = \frac{1}{(2\pi\sigma^2)^{1/2}} e^{-\frac{(y_i - \mu_i)^2}{2\sigma^2}}.$$

Assim, a função de verosimilhança vem

$$L = \prod_{i=1}^{n_0} (1 - F_i) \prod_{i=n_0+1}^{n_0+n_1} \frac{1}{(2\pi\sigma^2)^{1/2}} e^{-\frac{(y_i - \mu_i)^2}{2\sigma^2}},$$

onde o primeiro produto é feito sobre as observações de $y_i = 0$ e o segundo produto sobre as n_1 observações em que $y_i > 0$. A função de log-verosimilhança é

$$\ln L = \sum_{i=1}^{n_0} \ln(1 - F_i) - \frac{n_1}{2}\ln(2\pi) - \frac{n_1}{2}\ln(\sigma^2) - \sum_{i=n_0+1}^{n_0+n_1} \frac{(y_i - \mu_i)^2}{2\sigma^2}. \tag{11}$$

Para maximizar esta função de log-verosimilhança pode-se usar o algoritmo de Newton-Raphson. A expressão do estimador fornecido por este algoritmo é

$$\theta^{(m)} = \theta^{(m-1)} - \left[\frac{\partial^2 \ln L}{\partial\theta\partial\theta'}\right]^{-1}_{\theta=\theta^{(m-1)}} \cdot \left[\frac{\partial \ln L}{\partial\theta}\right]_{\theta=\theta^{(m-1)}},$$

onde $\theta = \begin{bmatrix} \beta_1 & \beta_2 & \ldots & \beta_k & \sigma^2 \end{bmatrix}^T$ é um vector de dimensão $(k+1)$. A matriz $\left[\frac{\partial^2 \ln L}{\partial\theta\partial\theta'}\right]$ corresponde à Hessiana da função de log-verosimilhança.

Uma abordagem alternativa consiste em utilizar o algoritmo dos *scores* de Fisher. Enquanto que o método de Newton-Raphson usa a inversa da matriz Hessiana da função de log-verosimilhança, o método dos *scores* de Fisher usa a inversa do valor esperado da Hessiana, o que corresponde à negativa da inversa da matriz de informação de Fisher. A utilização do método dos *scores* simplifica consideravelmente as expressões envolvidas. Neste último caso tem-se

$$\begin{aligned}\boldsymbol{\theta}^{(m)} &= \boldsymbol{\theta}^{(m-1)} - \left[E\left[\frac{\partial^2 \ln L}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right]\right]^{-1}_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(m-1)}} \cdot \left[\frac{\partial \ln L}{\partial\boldsymbol{\theta}}\right]_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(m-1)}} \\ &= \boldsymbol{\theta}^{(m-1)} + \left[I\left(\boldsymbol{\theta}\right)\right]^{-1}_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(m-1)}} \cdot \left[\frac{\partial \ln L}{\partial\boldsymbol{\theta}}\right]_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(m-1)}}.\end{aligned}$$

As derivadas de 1^a ordem da função de log-verosimilhança vêm

$$\begin{aligned}\frac{\partial \ln L}{\partial\boldsymbol{\beta}} &= -\frac{1}{\sigma}\sum_{i=1}^{n_0}\frac{f_i\mathbf{x}_i}{1-F_i} + \frac{1}{\sigma^2}\sum_{i=n_0+1}^{n_0+n_1}\left(y_i - \mathbf{x}_i^T\boldsymbol{\beta}\right)\mathbf{x}_i \\ \frac{\partial \ln L}{\partial\sigma^2} &= \frac{1}{2\sigma^3}\sum_{i=1}^{n_0}\frac{\left(\mathbf{x}_i^T\beta\right)f_i}{1-F_i} - \frac{N_1}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=n_0+1}^{n_0+n_1}\left(y_i - \mathbf{x}_i^T\boldsymbol{\beta}\right)^2,\end{aligned}$$

onde $F_i = \Phi\left(\frac{\mu_i}{\sigma}\right)$, $f_i = \phi\left(\frac{\mu_i}{\sigma}\right)$, $\boldsymbol{\beta} = \begin{bmatrix} \beta_1 & \beta_2 & \ldots & \beta_k \end{bmatrix}^T$ e $\mathbf{x}_i = \begin{bmatrix} x_{i1} & x_{i2} & \ldots & x_{ik} \end{bmatrix}^T$.

A matriz de informação é apresentada em Amemiya (1973) e é dada por

$$I\left(\boldsymbol{\theta}\right) = \begin{bmatrix} \sum_{i=1}^{n_0+n_1} a_i\mathbf{x}_i\mathbf{x}_i^T & \sum_{i=1}^{n_0+n_1} b_i\mathbf{x}_i \\ \sum_{i=1}^{n_0+n_1} b_i\mathbf{x}_i^T & \sum_{i=1}^{n_0+n_1} c_i \end{bmatrix},$$

onde

$$\begin{aligned}a_i &= -\frac{1}{\sigma^2}\left(\frac{\mu_i}{\sigma}f_i - \frac{f_i^2}{1-F_i} - F_i\right) \\ b_i &= \frac{1}{2\sigma^3}\left(\frac{\mu_i^2}{\sigma^2}f_i + f_i - \frac{\mu_i f_i^2}{\sigma\left(1-F_i\right)}\right) \\ c_i &= -\frac{1}{4\sigma^4}\left(\frac{\mu_i^3}{\sigma^3}f_i + \frac{\mu_i}{\sigma}f_i - \frac{\mu_i^2 f_i^2}{\sigma^2\left(1-F_i\right)} - 2F_i\right).\end{aligned}$$

Os estimadores de máxima verosimilhança para este modelo são consistentes e assimptoticamente normais com covariância assimptótica dada pela inversa da matriz de informação. Veja-se Amemiya (1973).

As expressões dos estimadores para amostras truncadas podem ser derivadas de forma semelhante às expressões para amostras censuradas.

Olsen (1978) mostrou que, efectuando uma reparametrização do modelo, pode-se garantir que a matriz de segundas derivadas da função de log-verosimilhança, em ordem aos parâmetros, é semidefinida negativa. Esse resultado implica que a função de verosimilhança tem um único máximo. Assim, independentemente dos valores de partida, desde que o processo iterativo continue até que se alcance uma solução, essa solução será o máximo global da função de verosimilhança. Segundo Maddala (1983), pode-se também provar que as funções de verosimilhança para modelos duplamente censurados e modelos truncados, após serem reparametrizados, também são globalmente côncavas.

A reparametrização, para a qual Olsen mostrou que a solução da maximização da função de log-verosimilhança é única, consiste em dividir a variável endógena pelo desvio padrão σ. Assim, a variável modelizada passa a ser

$$hY_i^* \sim N\left(\sum_{j=1}^{k} x_{ij}\beta_j^o, 1\right)$$

onde $h = \sigma^{-1}$ e $\beta_j^o = \frac{\beta_j}{\sigma}$. Após esta reparametrização, a função de log-verosimilhança pode-se escrever como

$$\ln L^o = \sum_{i=1}^{n_0} \ln\left(1 - F_i^o\right) - \frac{n_1}{2}\ln\left(2\pi\right) + n_1 \ln\left(h\right) - \sum_{i=n_0+1}^{n_0+n_1} \frac{\left(hy_i - \mu_i^o\right)^2}{2}$$

onde $\mu_i^o = E\left[hy_i^*\right]$. Após estimar o modelo reparametrizado, obtêm-se de forma imediata as estimativas dos parâmetros do modelo original, bastando calcular $\hat{\sigma} = \frac{1}{h}$ e multiplicar

as restantes estimativas por $\hat{\sigma}$.

3.3 A generalização dos modelos tobit

3.3.1 Funções de verosimilhança e estimação

Tal como se referiu na secção anterior, nos modelos de regressão tobit a variável aleatória Y_i^* tem distribuição normal de média μ_i, desvio padrão σ e é gerada pelo modelo

$$Y_i^* \sim N\left(\sum_{j=1}^{k} x_{ij}\beta_j, \sigma^2\right) \qquad i = 1, ..., n,$$

ou seja,

$$E[Y_i^*] = \mu_i = \sum_{j=1}^{k} x_{ij}\beta_j.$$

Desta formulação pode-se constatar que os modelos tobit apresentam duas fortes restrições: as observações são provenientes de uma distribuição normal e a média é combinação linear das variáveis explicativas. Para o presente trabalho é necessário abandonar estas duas hipóteses, uma vez que os fenómenos que se pretendem modelizar poderão seguir outras distribuições e, embora o valor esperado desses processos possa resultar de uma combinação linear das variáveis que condicionam o comportamento da distribuição, também se poderão verificar outros tipos de dependências funcionais.

A generalização dos modelos tobit será apresentada, um pouco à semelhança dos mo-

delos lineares generalizados, da seguinte forma

$$Y_i^* \sim f(y_i^*; \theta_i, \phi) \tag{12}$$

$$\mu_i = E[Y_i^*] = h(\theta_i) \tag{13}$$

$$\mu_i = g^{-1}\left(\sum_{j=1}^{k} x_{ij}\beta_j\right), \tag{14}$$

onde $f(y_i^*; \theta_i, \phi)$ é uma função densidade com parâmetros θ_i e ϕ. É frequente que estes parâmetros não correspondam aos parâmetros habituais da distribuição usada, dado que se reparametriza a função densidade por forma a que $E[Y_i^*]$ seja função apenas de θ_i.

Na expressão (12) pressupõe-se que a distribuição de Y_i^* é caracterizada por dois parâmetros. Os modelos podem ser utilizados com distribuições com um número diferente de parâmetros, mas as distribuições com dois parâmetros abrangem os casos de maior interesse, nomeadamente para os fenómenos que serão analisados neste trabalho. Considera-se que apenas um dos parâmetros da distribuição é variável com as características exógenas, enquanto que o outro é fixo.

A função $h(\cdot)$ estabelece a relação entre o valor esperado e o parâmetro θ_i, enquanto a função $g(\cdot)$ é monótona e diferenciável e é usualmente designada de função de ligação.

A distribuição normal tratada nos modelos tobit, é um caso particular da formalização apresentada em (12), onde $Y_i^* \sim N(\theta_i, \phi)$, com $\theta_i = \mu_i$ e $\phi = \sigma^2$, com $h(\theta_i) = \theta_i$ e com $g(\cdot)$ função identidade.

A função de verosimilhança de uma amostra de dimensão n, $(y_1, y_2, ..., y_n)$, censurada à esquerda de uma constante c será

$$L = \prod_{i=1}^{n_0} \int_{-\infty}^{c} f(y; \theta_i, \phi)\, dy \cdot \prod_{i=n_0+1}^{n_0+n_1} f(y_i; \theta_i, \phi) \qquad y_i \geqslant c, \tag{15}$$

onde $n_0 + n_1 = n$. A primeira parcela de (15) refere-se às n_0 observações da amostra em

que se observou o valor c (observações censuradas) enquanto a segunda parcela refere-se às restantes observações.

Caso as n observações resultem de uma amostra truncada à esquerda de c, então a função de verosimilhança será

$$L = \prod_{i=1}^{n} \frac{f(y_i; \theta_i, \phi)}{\int_c^{+\infty} f(y; \theta_i, \phi)\, dy} \qquad y_i \geqslant c.$$

Os casos em que a censura ou a truncagem se verificam à direita são modelizados de forma análoga.

Se o processo de amostragem for mais complexo, por exemplo, truncado para valores inferiores a uma constante c_1 e censurado para valores superiores a uma constante c_2, com $c_1 < c_2$, então a sua função de verosimilhança será

$$L = \prod_{i=1}^{n_0} \frac{\int_{c_2}^{+\infty} f(y; \theta_i, \phi)\, dy}{\int_{c_1}^{+\infty} f(y; \theta_i, \phi)\, dy} \cdot \prod_{i=n_0+1}^{n_0+n_1} \frac{f(y_i; \theta_i, \phi)}{\int_{c_1}^{+\infty} f(y; \theta_i, \phi)\, dy},$$

onde, mais uma vez, $n_0 + n_1 = n$ e a primeira parcela é o produto referente às n_0 observações censuradas da amostra, enquanto a segunda se refere às observações não censuradas.

No caso de a amostra estar truncada para valores fora do intervalo $[c_1; c_2]$ então a função de verosimilhança será

$$L = \prod_{i=1}^{n} \frac{f(y_i; \theta_i, \phi)}{\int_{c_1}^{c_2} f(y; \theta_i, \phi)\, dy}.$$

Tal como os modelos de regressão tobit, a estimação destes modelos poderá ser efectuada recorrendo aos algoritmos de Newton-Raphson ou dos *scores* de Fisher. Dependendo do tipo de restrições no processo de recolha das observações, da distribuição da variável Y_i^* e da expressão de $g(\cdot)$, têm-se diferentes expressões para $\frac{\partial \ln L}{\partial \beta}$, para a matriz Hessiana e para

a matriz de informação $I(\boldsymbol{\beta})$. Quando a variável Y_i^* não segue uma distribuição normal e/ou $g(\cdot)$ não é a função identidade, as expressões das primeiras derivadas da função de log-verosimilhança e da matriz de informação ficam, regra geral, bastante complexas. Não sendo objectivo deste trabalho a derivação dessas expressões, elas não serão indicadas para nenhum dos modelos à frente apresentados. Para a maximização das funções de log-verosimilhança recorreu-se à utilização de um *software* que permite a estimação pelo método de Newton-Raphson, neste caso, o TSP (Time Series Processor) na versão 4.4.

3.3.2 Inferência sobre os parâmetros

Para proceder a inferências sobre os parâmetros estimados para os modelos em estudo utiliza-se o teste da razão de verosimilhanças. Na exposição que se segue, considera-se a designação *modelo encaixado* como tendo o mesmo significado que foi atribuido no âmbito dos modelos lineares generalizados.

Considerem-se dois modelos encaixados, com número fixo de parâmetros, e que $\hat{\boldsymbol{\theta}}$ e $\hat{\boldsymbol{\theta}}^*$ são os vectores das estimativas dos parâmetros dos modelos, sem e com restrições lineares sobre os parâmetros, respectivamente. Sejam $L\left(\hat{\boldsymbol{\theta}};y\right)$ e $L\left(\hat{\boldsymbol{\theta}}^*;y\right)$ o valor da verosimilhança obtida com os dois vectores de estimativas dos parâmetros, então a razão de verosimilhanças é

$$\lambda = \frac{L\left(\hat{\boldsymbol{\theta}}^*;y\right)}{L\left(\hat{\boldsymbol{\theta}};y\right)}.$$

Pode-se mostrar que, se as hipóteses associadas às restrições lineares forem verdadeiras, então

$$LR = -2\ln\lambda = -2\left[l\left(\hat{\boldsymbol{\theta}}^*;y\right) - l\left(\hat{\boldsymbol{\theta}};y\right)\right] = 2\left[l\left(\hat{\boldsymbol{\theta}};y\right) - l\left(\hat{\boldsymbol{\theta}}^*;y\right)\right]$$

é assimptoticamente distribuído como uma variável aleatória qui-quadrado com d graus de liberdade, sendo d igual ao número de restrições lineares (linearmente independentes) impostas sobre o vector dos parâmetros. O procedimento para realizar os testes sobre os parâmetros é idêntico ao que foi descrito para os modelos lineares generalizados, residindo a única diferença no facto de agora se calcular a estatística de teste directamente através da função de log-verosimilhança.

À semelhança dos modelos lineares generalizados, a inferência sobre um parâmetro individual poderá ser efectuada através da comparação de modelos encaixados, ou recorrendo à distribuição assimptótica do estimador de máxima verosimilhança

$$\hat{\theta}_j \dot{\sim} N\left(\theta_j, \nu_{jj}\right),$$

onde ν_{jj} é o elemento (j,j) da inversa da matriz de informação de Fisher.

3.4 Alguns exemplos de modelos tobit generalizados

Uma breve análise da apresentação efectuada na secção anterior, permite constatar que os modelos tobit generalizados facultam a formalização de inúmeros modelos, pois podem-se estabelecer diversas combinações de funções de distribuição e de funções de ligação. Nos pontos seguintes são exemplificados alguns casos de modelizações que se enquadram na generalização dos modelos de regressão tobit e que são de particular interesse para o presente estudo.

3.4.1 A modelização com distribuição lognormal

Definição e algumas propriedades da distribuição lognormal A utilização de modelos tipo tobit com função de distribuição lognormal corresponde a uma pequena generalização dos modelos tobit, pois, regra geral, apesar da distribuição utilizada não

ser a normal, as propriedades da distribuição lognormal levam a que a sua estimação seja semelhante à da normal.

Genericamente, diz-se que Y tem *distribuição lognormal* quando $X = \ln(Y)$ tem distribuição normal, com Y podendo assumir qualquer valor superior a 0. Uma análise detalhada da distribuição lognormal pode ser encontrada em Johnson, Kotz e Balakrishnan (1999). A função densidade de probabilidade de Y é

$$f(y;\mu,\sigma) = \frac{1}{y\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{[\ln(y)-\mu]^2}{2\sigma^2}\right\} \qquad y>0 \quad (\sigma>0). \tag{16}$$

O r-ésimo momento de Y em torno de zero é dado por

$$\mu'_r = E[Y^r] = \exp\left(r\mu + \frac{1}{2}r^2\sigma^2\right),$$

de onde resulta que

$$E[Y] = \exp\left(\mu + \frac{\sigma^2}{2}\right) \qquad \text{e} \qquad Var[Y] = \exp(2\mu+\sigma^2)\left[\exp(\sigma^2)-1\right].$$

O terceiro momento central é

$$\mu_3[Y] = \exp(\sigma^2)^{\frac{3}{2}}\left[\exp(\sigma^2)-1\right]^2\left[\exp(\sigma^2)+2\right]\exp(3\mu).$$

Note-se que, sendo $X = \ln(Y)$, $E(X) = \mu$ e $var(X) = \sigma^2$.

De referir também que a moda de Y é $e^{\mu-\sigma^2}$, e que a sua mediana é e^{μ}.

A função de distribuição da lognormal, dadas as suas relações com a distribuição normal, é $\Phi\left(\frac{\ln(y)-\mu}{\sigma}\right)$.

No âmbito dos modelos em estudo, em que os valores observados estão sujeitos a truncagem e/ou censura, o valor esperado $E[\min(Y,c_2)|Y>c_1]$, assume particular inte-

resse. Guiahi (2000) apresenta a expressão para este valor esperado:

$$E\left[\min(Y,c_2)|Y>c_1\right] = \frac{e^{\mu+\frac{\sigma^2}{2}}\left[\Phi\left(\frac{\ln(c_2)-\mu-\sigma^2}{\sigma}\right)-\Phi\left(\frac{\ln(c_1)-\mu-\sigma^2}{\sigma}\right)\right]}{1-\Phi\left(\frac{\ln(c_1-\mu)}{\sigma}\right)} + \frac{c_2\left[1-\Phi\left(\frac{\ln(c_2)-\mu}{\sigma}\right)\right]}{1-\Phi\left(\frac{\ln(c_1-\mu)}{\sigma}\right)}.$$

Funções de verosimilhança da distribuição lognormal Quando a variável cujos parâmetros se pretendem estimar tem distribuição lognormal, é usual estimarem-se os parâmetros de $X=\ln(Y)$, que segue uma distribuição normal. Em resultado desta transformação da variável alcatória, ao contrário do que é habitual quando se utilizam outras distribuições, não se modeliza o valor esperado condicionado de Y, mas sim de $\ln(Y)$. Assim, o modelo tobit generalizado com distribuição lognormal é estimado quase como o modelo tobit original, podendo as únicas diferenças residir na forma como o parâmetro μ é condicionado pelas variáveis explicativas. Caso esse parâmetro seja função linear das variáveis explicativas, as funções de log-verosimilhança têm as mesmas expressões.

No caso de a amostra de Y ser truncada para valores das observações fora do intervalo $[c_1,c_2]$, a função de log-verosimilhança de X é

$$l=\sum_{i=1}^{n}\left\{-\frac{1}{2}\ln\left(2\pi\sigma^2\right)-\frac{(x_i-\mu_i)^2}{2\sigma^2}-\ln\left[\Phi\left(\frac{\ln(c_2)-\mu_i}{\sigma}\right)-\Phi\left(\frac{\ln(c_1)-\mu_i}{\sigma}\right)\right]\right\}.$$

Os casos em que a truncagem se verifica só para valores inferiores a c_1 ou só superiores a c_2, podem ser facilmente deduzidos da expressão anterior, bastando substituir $\Phi\left(\frac{c_2-\mu_i}{\sigma}\right)$ por 1 ou $\Phi\left(\frac{c_1-\mu_i}{\sigma}\right)$ por zero, respectivamente.

Se a amostra for truncada para valores inferiores a c_1 e censurada para valores supe-

riores a c_2, a função de log-verosimilhança será

$$l = \sum_{i=1}^{n_0} \left[\ln\left(1 - \Phi\left(\frac{\ln(c_2) - \mu_i}{\sigma}\right)\right) - \ln\left(1 - \Phi\left(\frac{\ln(c_1) - \mu_i}{\sigma}\right)\right)\right] + \sum_{i=n_0+1}^{n_0+n_1} \left[-\frac{1}{2}\ln\left(2\pi\sigma^2\right) - \frac{(x_i - \mu_i)^2}{2\sigma^2} - \ln\left(1 - \Phi\left(\frac{\ln(c_1) - \mu_i}{\sigma}\right)\right)\right].$$

Estas expressões baseiam-se na hipótese de que apenas o parâmetro μ, da distribuição de cada observação, é condicionada pelo valor assumido pelas respectivas variáveis explicativas, assumindo-se que o parâmetro σ é fixo para todas as observações.

Como, neste caso, a estimação dos parâmetros é efectuada recorrendo à transformação de variável $X = \ln(Y)$, que tem distribuição normal, as propriedades dos estimadores de máxima verosimilhança dos modelos tobit também são válidas para os modelos generalizados com distribuição lognormal, desde que a função de ligação utilizada seja a identidade, ou seja, $\mu_i = \sum_j x_{ij}\beta_j$.

Resultados da aplicação do modelo com distribuição lognormal a dados simulados Apesar de já existirem diversos resultados sobre a qualidade das estimativas dos modelos tobit, com o objectivo de ilustrar o comportamento dos modelos tobit generalizados na estimação de parâmetros de distribuições lognormal, recorreu-se à aplicação dos vários modelos a dados simulados. Assim, foi gerada aleatoriamente uma amostra de 10000 observações de variáveis aleatórias Y_i com distribuição lognormal, considerando que $\sigma = 1$ e que

$$\mu_i = \beta_1 + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \beta_4 x_{i,4} + \beta_5 x_{i,5}$$

com

$$\beta_1 = -2 \quad \beta_2 = 0,5 \quad \beta_3 = 0,3 \quad \beta_4 = -0,3 \quad \beta_5 = -0,5$$

e onde x_{ij} são variáveis binárias, ou seja

$$x_{ij} = \begin{cases} 1 & \text{se a observação } i \text{ tem a característica } j \\ 0 & \text{caso contrário} \end{cases}.$$

Os valores assumidos pelas 4 variáveis explicativas foram também gerados aleatoriamente, através de distribuições de Bernoulli. Na geração das variáveis explicativas, assumiu-se que $\Pr(x_2 = 1) = 0,5$, $\Pr(x_3 = 1) = 0,75$, $\Pr(x_4 = 1) = 0,25$ e $\Pr(x_5 = 1) = 0,6$. Assumiu-se também que não existia qualquer correlação entre essas variáveis.

Note-se que, por a modelização ser feita para a variável $X_i = \ln(Y_i)$, se está a considerar uma função de ligação identidade, pois $\mu_i = \sum_{j=1}^{5} x_{ij}\beta_j$, assumindo que $x_{i1} = 1$, $(i = 1, ..., n)$. No entanto, o valor esperado de Y_i é $\exp\left(\mu_i + \frac{\sigma^2}{2}\right)$, pelo que temos

$$\begin{aligned} E[Y_i] &= \exp\left[\beta_1 + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \beta_4 x_{i,4} + \beta_5 x_{i,5} + \frac{\sigma^2}{2}\right] \\ &= \exp\left[\beta_1' + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \beta_4 x_{i,4} + \beta_5 x_{i,5}\right], \end{aligned}$$

onde $\beta_1' = \beta_1 + \frac{\sigma^2}{2}$, o qual é um resultado semelhante ao que se obtém quando, para outras distribuições, se modeliza directamente o valor esperado com uma função de ligação logarítmica.

Após a geração das observações, estimaram-se os parâmetros do modelo sem restrições e com diversas combinações de censuras (para valores superiores a uma constante) e truncagens (para valores inferiores a uma constante) em diversos valores, para exemplificar a robustez do método de estimação a vários níveis de restrições. Com a construção deste exemplo não se pretende avaliar a qualidade das estimativas resultantes da aplicação destes modelos, caso em que seria necessário proceder a um estudo de simulação, sendo o único objectivo ilustrar a alteração das estimativas face a diferentes tipos de restrições nas

	Truncagem	Censura					
		S/ Censura	1,00	0,60	0,40	0,20	0,15
β_1	S/ Truncagem	-1,980	-1,986	-1,988	-1,985	-1,981	-1,979
β_2		0,511	0,512	0,513	0,514	0,502	0,504
β_3		0,311	0,313	0,311	0,309	0,302	0,303
β_4		-0,317	-0,315	-0,315	-0,314	-0,322	-0,322
β_5		-0,559	-0,555	-0,552	-0,553	-0,545	-0,549
σ		1,001	1,001	0,998	0,999	0,994	0,996
β_1	0,01	-1,980	-1,986	-1,987	-1,984	-1,978	-1,973
β_2		0,516	0,517	0,518	0,520	0,508	0,512
β_3		0,306	0,308	0,306	0,304	0,296	0,297
β_4		-0,312	-0,310	-0,309	-0,308	-0,317	-0,317
β_5		-0,566	-0,562	-0,559	-0,561	-0,554	-0,560
σ		1,006	1,006	1,003	1,006	1,001	1,007
β_1	0,03	-1,985	-1,991	-1,991	-1,989	-1,976	-1,975
β_2		0,515	0,515	0,515	0,518	0,498	0,504
β_3		0,312	0,316	0,312	0,309	0,296	0,299
β_4		-0,305	-0,301	-0,300	-0,299	-0,307	-0,308
β_5		-0,564	-0,559	-0,553	-0,557	-0,542	-0,551
σ		1,004	1,005	0,999	1,003	0,990	0,999
β_1	0,05	-2,000	-2,009	-2,004	-2,006	-1,977	-1,985
β_2		0,499	0,500	0,499	0,504	0,467	0,476
β_3		0,332	0,337	0,332	0,330	0,312	0,322
β_4		-0,311	-0,306	-0,304	-0,303	-0,312	-0,318
β_5		-0,558	-0,552	-0,544	-0,550	-0,522	-0,537
σ		1,009	1,010	1,002	1,009	0,987	1,006
β_1	0,10	-2,045	-2,068	-2,045	-2,007	-1,922	-2,025
β_2		0,524	0,528	0,523	0,541	0,428	0,479
β_3		0,372	0,384	0,373	0,376	0,321	0,392
β_4		-0,287	-0,279	-0,274	-0,270	-0,259	-0,272
β_5		-0,582	-0,573	-0,556	-0,574	0,477	-0,543
σ		1,011	1,015	0,998	1,016	0,914	0,989

Figura 2: Resultados de simulação com distribuição lognormal

observações.

A Figura 2 contém os resultados da estimação dos diversos modelos com distribuição lognormal.

A quantidade de observações afectadas pelas restrições impostas à informação utilizada na estimação encontram-se na Figura 3.

Os resultados da estimação permitem constatar uma reduzida sensibilidade do valor das estimativas à quantidade de informação sujeita a truncagem ou censura. Verifica-se que, mesmo no modelo com mais informação restringida (censura em 0,15 e truncagem em 0,10, caso em que 3572 observações são truncadas e 4983 são censuradas), as estimativas estão muito próximas daquelas que resultam do modelo sem qualquer restrição. No entanto, constata-se, neste conjunto de dados simulados, que em todos as combinações

Truncagem				
Valor de Truncagem	0,01	0,03	0,05	0,10
Nº Observ. Truncadas	52	710	1578	3572

Censura					
Valor de Censura	0,15	0,20	0,40	0,60	1,00
Nº Observ. Censuradas	4983	3911	1844	999	426

Figura 3: Nºde observações censuradas e truncadas - distribuição lognormal

de restrições sobre a informação, o parâmetro β_5 está com um enviesamento que ronda -0,05. A dificuldade na estimação deste parâmetro, deve estar associada ao facto de as observações com $x_5 = 1$ terem um menor valor esperado, e por isso, estarem mais sujeitas ao efeito da truncagem.

Procedeu-se à estimação de parâmetros de outros conjuntos de dados simulados, com as mesmas características, tendo-se sempre obtido resultados que conduzem a conclusões semelhantes.

3.4.2 A modelização com distribuição gama

Definição e algumas propriedades da distribuição gama A variável aleatória Y tem uma *distribuição gama* se a sua função densidade for dada por

$$f(y; m, \alpha) = \frac{\alpha^m}{\Gamma(m)} e^{-\alpha y} y^{m-1} \qquad y > 0 \qquad (m > 0; \alpha > 0), \tag{17}$$

onde $\Gamma(m)$ corresponde à função gama, ou seja

$$\Gamma(m) = \int_0^{+\infty} t^{m-1} e^{-t} dt.$$

Diversas distribuições podem ser derivadas da função densidade gama, como por exemplo, a sua forma standard, que é obtida impondo que $\alpha = 1$, ou a *distribuição exponencial*,

a qual se obtém quando $m = 1$. Quando m é um inteiro positivo tem-se a *distribuição de Erlang*. Mais detalhes sobre a distribuição gama podem ser encontrados, por exemplo, em Johnson, Kotz e Balakrishnan (1999).

A função de distribuição gama é dada por

$$\Pr[Y \leqslant y] = \frac{\alpha^m}{\Gamma(m)} \int_0^y e^{-\alpha t} t^{m-1} dt,$$

e, fazendo uma mudança de variável $z = \alpha t$, tem-se

$$\Pr[Y \leqslant y] = \Gamma(m)^{-1} \int_0^{\alpha y} e^{-z} z^{m-1} dz = \frac{\Gamma(y\alpha, m)}{\Gamma(m)},$$

onde

$$\Gamma(x, m) = \int_0^x t^{m-1} e^{-t} dt$$

é usualmente denominada de *função gama incompleta*. Por vezes utiliza-se a designação *função gama incompleta* para identificar a quantidade $\Pr[Y \leqslant y]$, como sucede, por exemplo, em Hogg e Klugman (1984).

A função de distribuição gama tem função geradora de momentos

$$M_Y(s) = \left(\frac{\alpha}{\alpha - s}\right)^m \qquad s < \alpha,$$

de onde resulta que

$$E(Y) = \frac{m}{\alpha}, \quad Var(Y) = \frac{m}{\alpha^2} = \frac{1}{m} E(Y)^2, \quad \mu_3(Y) = \frac{2m}{\alpha^3} = \frac{2}{m^2} E(Y)^3.$$

Seja $f(y; m, \alpha)$ a função densidade e $F(y; m, \alpha)$ a função de distribuição de Y, com

distribuição gama de parâmetros m e α. Assim, tem-se que

$$\begin{aligned} E\left[\min(Y,c_2)|Y>c_1\right] &= \frac{\left(\int_{c_1}^{c_2} yf\left(y;m,\alpha\right)dy + c_2\left[1-F\left(c_2;m,\alpha\right)\right]\right)}{\left[1-F\left(c_1;m,\alpha\right)\right]} \\ &= \frac{\frac{m}{\alpha}\left[F\left(c_2;m+1,\alpha\right)-F\left(c_1;m+1,\alpha\right)\right]}{\left[1-F\left(c_1;m,\alpha\right)\right]} \\ &\quad + \frac{c_2\left[1-F\left(c_2;m,\alpha\right)\right]}{\left[1-F\left(c_1;m,\alpha\right)\right]}. \end{aligned}$$

A função densidade de probabilidade da distribuição gama pode também ser apresentada na forma das distribuições de família de dispersão exponencial, isto é

$$f\left(y;m,\alpha\right) = \exp\left[\frac{-\frac{\alpha}{m}y + \ln\left(\frac{\alpha}{m}\right)}{\frac{1}{m}} + m\ln\left(my\right) - \ln y - \ln\Gamma\left(m\right)\right].$$

Ao reparametrizar a f.d.p. em função do seu valor esperado, $\mu = \frac{m}{\alpha}$, e de m, tem-se

$$f\left(y;\mu,\alpha\right) = \exp\left[m\left(-\mu^{-1}y - \ln\left(\mu\right)\right) + m\ln\left(my\right) - \ln y - \ln\Gamma\left(m\right)\right].$$

Funções de verosimilhança da distribuição gama A função de verosimilhança para a distribuição gama com observações truncadas à esquerda de uma constante c é

$$\begin{aligned} L &= \prod_{i=1}^{n} \frac{\exp\left[m\left(-\mu_i^{-1}y_i - \ln\left(\mu_i\right)\right) + m\ln\left(my_i\right) - \ln y_i - \ln\Gamma\left(m\right)\right]}{\int_c^{+\infty}\exp\left[m\left(-\mu_i^{-1}y - \ln\left(\mu_i\right)\right) + m\ln\left(my\right) - \ln y - \ln\Gamma\left(m\right)\right]dy} \\ &= \prod_{i=1}^{n} \frac{\exp\left[m\left(-\mu_i^{-1}y_i - \ln\left(\mu_i\right)\right) + m\ln\left(my_i\right) - \ln y_i - \ln\Gamma\left(m\right)\right]}{1-\frac{\Gamma\left(c\frac{m}{\mu_i},m\right)}{\Gamma(m)}} \end{aligned}$$

e a função de log-verosimilhança vem

$$\begin{aligned} l &= \ln L = \\ &= \sum_{i=1}^{n}\left(m\left(-\mu_i^{-1}y_i - \ln\left(\mu_i\right)\right) + m\ln\left(my_i\right) - \ln y_i - \ln\Gamma\left(m\right) - \ln\left(1-\frac{\Gamma\left(c\frac{m}{\mu_i},m\right)}{\Gamma\left(m\right)}\right)\right). \end{aligned}$$

As expressões anteriores pressupõem que, dos parâmetros da distribuição gama, apenas a média é condicionada pelas variáveis explicativas, assumindo-se que o parâmetro m é constante para todas as observações. Estes pressupostos implicam que a variância é proporcional a μ_i^2. Nas expressões anteriores não se assume qual a função de ligação que relaciona μ_i com as observações das variáveis explicativas x_{ij}.

No caso de a amostra ser censurada à esquerda de c, a função de log-verosimilhança é

$$\begin{aligned} l &= \sum_{i=1}^{n_0} \ln\left(\frac{\Gamma\left(c\frac{m}{\mu_i}, m\right)}{\Gamma(m)}\right) \\ &\quad + \sum_{i=n_0+1}^{n_0+n_1} \left(m\left(-\mu_i^{-1}y_i - \ln(\mu_i)\right) + m\ln(my_i) - \ln y_i - \ln\Gamma(m)\right). \end{aligned}$$

Ao contrário do que sucede para os modelos com distribuição lognormal, não está provado para os modelos gama que as diversas funções de verosimilhança sejam globalmente côncavas. Assim, não há garantia de que o modelo converge para o máximo global da função de verosimilhança, podendo eventualmente convergir para um máximo local.

Dado que se sabe que o resultado do modelo lognormal corresponde a um máximo global, sugere-se que, quando se pretenda estimar um modelo gama, se estime previamente um modelo lognormal, com função de ligação identidade e o mesmo tipo de restrições, e que se usem as estimativas desse modelo para definir os valores iniciais para o processo de estimação do modelo gama. No caso de a função de ligação do modelo gama ser logarítmica, identificando $\hat{\beta}_1, \hat{\beta}_2 ... \hat{\beta}_k, \hat{\sigma}$ como as estimativas de máxima verosimilhança do modelo lognormal e considerando os dois primeiros momentos das distribuições gama e

lognormal, têm-se como valores iniciais do modelo gama

$$\begin{aligned} \beta_1^0 &= \hat{\beta}_1 + \frac{1}{2}\hat{\sigma}^2 \\ \beta_j^0 &= \hat{\beta}_j \qquad j = 2, ..., k \\ m^0 &= \left(e^{\hat{\sigma}^2} - 1\right)^{-1}, \end{aligned}$$

onde o índice 1 identifica os termos independentes.

Aproximação numérica para a função de distribuição da gama No âmbito destes modelos, a estimação dos parâmetros da distribuição gama, recorrendo ao software TSP, apresenta uma dificuldade adicional: ao contrário do que acontece com a distribuição normal, não está definida (no TSP) uma aproximação da função de distribuição gama, ou da função gama incompleta que, como se viu atrás, permite, em conjunto com a função gama, calcular a densidade acumulada da distribuição gama. Devido a esta limitação do *software* utilizado, foi necessário implementar uma aproximação da função de distribuição gama. De seguida faz-se uma pequena apresentação e discussão das aproximações utilizadas.

Em Johnson, Kotz & Balakrishnan (1999) são apresentadas diversas aproximações e algoritmos para as funções de distribuição gama e qui-quadrado. Para aplicação neste trabalho utilizaram-se duas aproximações à função de distribuição gama que foram desenvolvidas para a distribuição do qui-quadrado. Ambas recorrem à função de distribuição normal. A primeira, de Wilson-Hilferty (1931), tem um expressão relativamente simples:

$$F_{\chi_v^2}(y) \simeq \Phi\left(\sqrt{\frac{9v}{2}}\left\{\left(\frac{y}{v}\right)^{\frac{1}{3}} - 1 + \frac{2}{9v}\right\}\right).$$

Dadas as relações entre as distribuições gama e qui-quadrado resulta que

$$\begin{aligned} F_{\gamma(m,\alpha)}(y) &= \frac{\Gamma(y\alpha, m)}{\Gamma(m)} \simeq \Phi\left(\sqrt{\frac{9\times 2m}{2}}\left\{\left(\frac{2\alpha y}{2m}\right)^{\frac{1}{3}} - 1 + \frac{2}{9\times 2m}\right\}\right) \\ &= \Phi\left(\sqrt{9m}\left\{\left(\frac{\alpha y}{m}\right)^{\frac{1}{3}} - 1 + (9m)^{-1}\right\}\right). \end{aligned}$$

A aproximação desenvolvida por Peizer e Pratt (1968), apresenta uma expressão bastante mais complexa

$$F_{\chi^2_v}(y) \simeq \begin{cases} \Phi\left(-\frac{\frac{1}{3}+\frac{0,08}{v}}{\sqrt{2v-2}}\right) & \text{se } y = v-1 \\ \Phi\left(\frac{y-v+\frac{2}{3}-\frac{0,08}{v}}{|y-(v-1)|}\left[(v-1)\ln\left(\frac{v-1}{y}\right)+y-(v-1)\right]^{1/2}\right) & \text{se } x \neq v-1 \end{cases},$$

ou seja

$$F_{\gamma(m,\alpha)}(y) \simeq \begin{cases} \Phi\left(-\frac{\frac{1}{3}+\frac{0,04}{m}}{\sqrt{4m-2}}\right) & \text{se } y = \frac{m}{\alpha} - \frac{1}{2\alpha} \\ \Phi\left(\frac{2m\left(y\frac{\alpha}{m}-1\right)+\frac{2}{3}-\frac{0,04}{m}}{\left|2m\left(y\frac{\alpha}{m}-1\right)+1\right|}\left[(2m-1)\ln\left(\frac{2m-1}{2\alpha y}\right)+2m\left(y\frac{\alpha}{m}-1\right)+1\right]^{1/2}\right) & \\ & \text{caso contrário} \end{cases}.$$

Para ilustrar a qualidade das aproximações fornecidas pelas expressões anteriores, calcularam-se em vários pontos, para distribuições gama com diferentes parâmetros, os desvios entre os valores resultantes das duas aproximações e aqueles que se obtêm através de procedimentos com elevada precisão (calculados através do software *Mathematica 3.0*). Essas diferenças são apresentadas na Figura 4.

Estes valores mostram que os resultados das aproximações são bastante satisfatórios, verificando-se, no entanto, que para os valores mais baixos do parâmetro m, quando se calcula a densidade acumulada para valores de x próximos de zero, o erro da aproximação

	x	0,01	0,05	0,1	0,3	0,5	1	2
m=0,6	Aprox. W-H	0,02493	0,00524	-0,00352	-0,00626	-0,00168	0,00358	0,00175
α=1,2	Aprox. P-P	-0,02687	-0,01649	-0,00374	0,01440	0,01641	0,01125	0,00351
m=0,8	Aprox. W-H	0,01779	0,00600	-0,00186	-0,00597	-0,00173	0,00294	0,00076
α=1,6	Aprox. P-P	-0,00520	-0,00431	-0,00097	0,00573	0,00676	0,00447	0,00110
m=1,2	Aprox. W-H	0,00811	0,00581	0,00056	-0,00487	-0,00137	0,00215	-0,00001
α=2,4	Aprox. P-P	-0,00039	-0,00073	-0,00027	0,00171	0,00222	0,00137	0,00021
m=2,5	Aprox. W-H	0,00042	0,00200	0,00189	-0,00260	-0,00063	0,00083	-0,00011
α=5	Aprox. P-P	0,00000	0,00001	0,00001	0,00017	0,00032	0,00017	0,00001

Figura 4: Erros das aproximações da função de distribuição gama

é significativo. Para os valores exemplificados não se verifica que uma aproximação seja sempre melhor que a outra, mas a precisão da aproximação de Peizer e Pratt para valores de m mais elevados é bastante superior, principalmente quando calculadas para valores de x próximos de 0.

Ling (1977, 1978) comparou a precisão de 4 aproximações para a distribuição do qui-quadrado, entre as quais, as aproximações de Wilson-Hilferty e a de Peizer e Pratt. A medida utilizada para avaliar a precisão foi o erro absoluto máximo gerado por cada uma das aproximações para diversos valores de x, para distribuições com graus de liberdade que variaram entre 5 e 240. Os resultados mostraram que os erros máximos da aproximação de Peizer e Pratt, para os casos analisados, foram sempre inferiores aos de Wilson-Hilferty.

Quer as simulações apresentadas na tabela anterior, quer o estudo de Ling, sugerem que a aproximação de Peizer e Pratt é mais precisa. No entanto, a expressão desta aproximação é mais complexa que a de Wilson-Hilferty, principalmente por ter três ramos (devido ao módulo da expressão para $y \neq \frac{m}{\alpha} - \frac{1}{2\alpha}$), o que poderá implicar maiores dificuldades computacionais para estimar os parâmetros dos modelos. Sempre que possível sugere-se a utilização da fórmula de Peizer e Pratt.

Dependendo do valor dos parâmetros envolvidos na estimação dos modelos e dos valores

para os quais se calcularem as densidades acumuladas, os erros das aproximações poderão provocar maiores ou menores enviesamentos na estimação dos parâmetros dos modelos de regressão tobit com distribuição gama.

Resultados da aplicação do modelo gama a dados simulados Tal como no caso dos modelos com distribuição lognormal, foram geradas aleatoriamente 10000 observações de variáveis aleatórias com distribuição gama de parâmetro $m = 1,2$ e com o valor esperado a ser condicionado pelas variáveis explicativas x_{ij} utilizadas no exemplo da lognormal. Definiu-se que o valor esperado da distribuição de cada observação y_i, para $i = 1, 2, ..., 10000$, seria

$$\mu_i = \exp\{\beta_1 + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \beta_4 x_{i,4} + \beta_5 x_{i,5}\},$$

com

$$\beta_1 = -1,5 \quad \beta_2 = 0,5 \quad \beta_3 = 0,3 \quad \beta_4 = -0,3 \quad \beta_5 = -0,5 \quad .$$

Neste exemplo a função de ligação é a logarítmica, ou seja, $g(\cdot) = \ln(\cdot)$.

O valor esperado das observações varia entre $\exp\{-1,5 - 0,3 - 0,5\} = 0,1003$ e $\exp\{-1,5 + 0,5 + 0,3\} = 0,4966$, referindo-se o 1^o valor a observações associadas a "casos" que possuam apenas as caracteristicas 1 e 2, e o 2^ovalor a "casos" que só possuam as características 3 e 4.

Geradas as observações, estimaram-se os parâmetros do modelo sem restrições e com combinações de censuras (à direita) e truncagens (à esquerda) em diversos valores, para ilustrar a robustez do método de estimação a vários níveis de restrições.

Primeiro estimaram-se os vários modelos utilizando a aproximação de Wilson-Hilferty para calcular a função de distribuição quer no ponto de censura, quer no ponto de truncagem. Os resultados apresentaram alguns desvios entre as estimativas e o valor

real dos coeficientes.

Para tentar melhorar a qualidade da estimação e dado que a aproximação de Peizer e Pratt é mais precisa, introduziu-se essa aproximação na estimação do modelo. No entanto, a sua utilização apresenta algumas dificuldades. Como a expressão da aproximação de Peizer e Pratt tem um módulo, a função não é diferenciável, o que impossibilita a utilização directa dessa aproximação na função de log-verosimilhança. No entanto, se apenas se calcular a função de distribuição gama para valores tal que $y < \frac{m}{\alpha} - \frac{1}{2\alpha}$, ou, $y > \frac{m}{\alpha} - \frac{1}{2\alpha}$, então a aproximação poderá ser utilizada, pois deixará de ser necessário definir o módulo. Uma vez que, quer a aproximação Peizer e Pratt, quer a de Wilson-Hilferty, aparentam maior precisão para valores bastante à direita do valor esperado da distribuição, os maiores desvios no cálculo da função de distribuição verificam-se na aba esquerda da distribuição. Assim, na definição da função de verosimilhança utilizou-se a aproximação de Peizer e Pratt para calcular a função de distribuição gama nos pontos de truncagem, mantendo-se a aproximação de Wilson-Hilferty nos pontos de censura. Como o menor valor de $\frac{m}{\alpha}$ utilizado na geração dos dados foi de $0,1003$, a truncagem só poderá ser aplicada para valores inferiores a $0,05851$. Os resultados da estimação dos vários modelos encontram-se na Figura 5.

As células no canto inferior direito do quadro não estão preenchidas porque na estimação do modelo, com truncagem para valores inferiores a 0,05 e censura para valores superiores a 0,15, não se conseguiu alcançar a convergência da função de log-verosimilhança. A célula a sombreado indica a única estimativa em que se rejeita a hipótese de o parâmetro ser igual ao coeficiente utilizado na geração dos dados, com um grau de confiança de 95%. Assim, da estimação dos parâmetros deste conjunto de observações resultaram, para todos os tipos de restrições, estimativas muito próximas dos verdadeiros valores dos coeficientes, mesmo quando a censura é feita para valores superiores a 0,15, o que representa a censura

	Truncagem	Censura					
		S/ Censura	1,00	0,60	0,40	0,20	0,15
β_1	S/ Truncagem	-1,484	-1,482	-1,483	-1,481	-1,469	-1,451
β_2		0,481	0,483	0,490	0,505	0,483	0,466
β_3		0,279	0,281	0,288	0,284	0,272	0,258
β_4		-0,285	-0,284	-0,291	-0,299	-0,316	-0,302
β_5		-0,493	-0,499	-0,498	-0,504	-0,493	-0,489
m		1,201	1,200	1,193	1,195	1,189	1,181
β_1	0,01	-1,483	-1,481	-1,483	-1,481	-1,473	-1,457
β_2		0,482	0,483	0,489	0,504	0,480	0,462
β_3		0,280	0,282	0,289	0,284	0,271	0,256
β_4		-0,285	-0,284	-0,291	-0,299	-0,301	-0,302
β_5		-0,495	-0,500	-0,500	-0,505	-0,493	-0,489
m		1,218	1,216	1,209	1,214	1,214	1,204
β_1	0,03	-1,482	-1,481	-1,483	-1,480	-1,463	-1,433
β_2		0,480	0,482	0,489	0,506	0,483	0,467
β_3		0,277	0,279	0,287	0,282	0,268	0,252
β_4		-0,282	-0,281	-0,289	-0,298	-0,302	-0,300
β_5		-0,497	-0,504	-0,504	-0,511	-0,504	-0,508
m		1,202	1,197	1,187	1,192	1,169	1,123
β_1	0,05	-1,489	-1,488	-1,493	-1,489	-1,470	
β_2		0,481	0,484	0,494	0,515	0,511	
β_3		0,283	0,286	0,296	0,292	0,291	
β_4		-0,288	-0,287	-0,297	-0,309	-0,332	
β_5		-0,497	-0,504	-0,506	-0,515	-0,528	
m		1,191	1,183	1,164	1,165	1,055	

Figura 5: Resultados de simulação com distribuição gama

de 5647 observações. Para este conjunto de dados, verificou-se que a utilização da aproximação de Peizer e Pratt, no cálculo da probabilidade acumulada no valor da truncagem, permitiu alcançar estimativas mais precisas do que quando só se utilizou a aproximação de Wilson-Hilferty.

Para que se possa ter noção da importância das restrições introduzidas, apresenta-se, na Figura 6, o número de observações restringidas em cada um dos pontos de truncagem e censura.

3.4.3 A modelização com distribuição inversa Gaussiana

Definição e algumas propriedades da distribuição inversa Gaussiana Uma variável aleatória Y tem distribuição inversa Gaussiana generalizada de parâmetros d, β e

Truncagem				
Valor de Truncagem	0,01	0,03	0,05	0,10
N° Observ. Truncadas	258	887	1560	3146

Censura					
Valor de Censura	0,15	0,20	0,40	0,60	1,00
N° Observ. Censuradas	5647	4589	2021	1041	269

Figura 6: N°de observações censuradas e truncadas - distribuição gama

ν, se tem como função densidade de probabilidade:

$$f(y;d,\beta,v) = \frac{1}{\sqrt{2\pi\beta y^3}} d e^{-(d-vy)^2/(2\beta y)} \qquad y>0 \quad (d>0,\beta>0,v>0). \tag{18}$$

Definindo $v = \frac{d}{\mu}$ e $\beta = \frac{d^2}{\lambda}$, a expressão anterior vem

$$f(y;\mu,\lambda) = \left[\frac{\lambda}{2\pi y^3}\right]^{\frac{1}{2}} \exp\left\{-\frac{\lambda(y-\mu)^2}{2\mu^2 y}\right\} \qquad y>0 \quad (\mu>0;\lambda>0). \tag{19}$$

A expressão (19) identifica a função densidade de probabilidade da distribuição inversa Gaussiana estandartizada e corresponde à parametrização clássica da inversa Gaussiana. Por conveniência, esta parametrização será designada por IG1. Quando em (19) se assume que $\mu = 1$, tem-se a forma estandartizada da distribuição de Wald. A distribuição IG1 tem como valor esperado e 2^oe 3^omomentos centrais:

$$E(Y) = \mu, \quad Var(Y) = \frac{\mu^3}{\lambda}, \quad \mu_3(Y) = \frac{3}{\lambda^2}\mu^5. \tag{20}$$

A distribuição inversa Gaussiana pode ainda ser escrita de outras formas alternativas, nomeadamente

$$f(y;\mu,\phi) = \left[\frac{\mu\phi}{2\pi y^3}\right]^{\frac{1}{2}} \exp\left\{-\frac{\phi y}{2\mu} + \phi - \frac{\mu\phi}{2y}\right\} \qquad y>0 \quad (\mu>0;\phi>0), \tag{21}$$

que se obtém definindo $\phi = \frac{\mu}{\lambda}$. Para distinguir esta parametrização da inversa Gaussiana daquela que foi introduzida em (19), a parametrização correspondente a (21) será designada por IG2. Os principais momentos da distribuição, com esta parametrização, são:

$$E(Y) = \mu, \quad Var(Y) = \frac{\mu^2}{\phi}, \quad \mu_3(Y) = \frac{3}{\phi^2}\mu^3. \tag{22}$$

Outras formas equivalentes de escrever a função densidade da inversa Gaussiana podem ser encontradas em Johnson, Kotz e Balakrishnan (1999).

A opção por uma das parametrizações, IG1 ou IG2, para proceder à modelização de variáveis aleatórias, não é indiferente, dado que a cada uma delas corresponde um diferente padrão de heterocedasticidade, como se depreende da comparação de (20) com (22). Na parametrização "clássica", a IG1, que corresponde à parametrização utilizada nos modelos lineares generalizados, assume-se que a variância é proporcional ao cubo do valor esperado. Na parametrização IG2, por seu lado, assume-se que a variância é proporcional ao quadrado do valor esperado.

Se compararmos os momentos centrais das distribuições gama e inversa Gaussiana na forma IG2, verificamos que em ambas as distribuições a variância é função do quadrado do valor esperado vezes um parâmetro (m na distribuição gama e ϕ na distribuição inversa gaussiana) e que assumindo a mesma média e variância para as duas distribuições, o coeficiente de assimetria da distribuição inversa gaussiana é sempre $\frac{3}{2}$ do da distribuição gama.

No que se segue da exposição será sempre utilizada a parametrização IG2. Note-se que, devido ao seu diferente padrão de heterocedasticidade, a modelização com a distribuição IG2 não é directamente comparável com a modelização com a distribuição inversa Gaussiana recorrendo aos modelos lineares generalizados.

Se Y tem distribuição inversa Gaussiana com parâmetros μ e ϕ, então $\frac{Y}{\mu}$ tem distribuição de Wald estandardizada com parâmetro ϕ, ou seja,

$$Y \sim ig(\mu, \phi) \Rightarrow \frac{Y}{\mu} \sim ig(1, \phi).$$

A função densidade acumulada de uma variável aleatória de Wald com parâmetro ϕ é (Johnson, Kotz e Balakrishnan (1999))

$$F(y;\phi) = \Phi\left((y-1)\sqrt{\frac{\phi}{y}}\right) + e^{2\phi}\Phi\left(-(y+1)\sqrt{\frac{\phi}{y}}\right),$$

onde $\Phi(\cdot)$ é a função de distribuição normal estandardizada. Dada a relação entre as distribuições inversa Gaussiana e de Wald, facilmente se conclui que a função de distribuição da inversa Gaussiana é

$$G(y;\mu,\phi) = \Phi\left((y-\mu)\sqrt{\frac{\phi}{\mu y}}\right) + e^{2\phi}\Phi\left(-(y+\mu)\sqrt{\frac{\phi}{\mu y}}\right).$$

Peter Ter Berg (1994) apresenta uma expressão para $E[X]$, com

$$X = \begin{cases} 0 & Y \leqslant c_1 \\ Y - c_1 & c_1 < Y \leqslant c_2 \\ c_2 - c_1 & c_2 < Y \end{cases},$$

em que Y segue uma distribuição inversa Gaussiana com parâmetros μ e ϕ. Assim,

$$\begin{aligned} E[X] &= (c_2 - c_1)[1 - G(c_2;\mu,\phi)] \\ &\quad + (\mu + c_1)[G(c_1;\mu,\phi) - G(c_2;\mu,\phi)] \\ &\quad + 2\mu\left\{\Phi\left[(\mu - c_1)\sqrt{\frac{\phi}{\mu c_1}}\right] - \Phi\left[(\mu - c_2)\sqrt{\frac{\phi}{\mu c_2}}\right]\right\}. \end{aligned}$$

Como

$$E\left[\min(Y, c_2)|Y > c_1\right] = \frac{E\left[X\right]}{1 - G\left(c_1; \mu, \phi\right)} + c_1$$

facilmente se passa da expressão de $E\left[X\right]$ fornecida por Berg (1994) para $E\left[\min(Y, c_2)|Y > c_1\right]$.

Funções de verosimilhança da distribuição inversa Gaussiana A função de log-verosimilhança para uma amostra com distribuição inversa Gaussiana é

$$l = \sum_{i=1}^{n} \left(-\frac{\phi y_i}{2\mu_i} + \phi - \frac{\phi \mu_i}{2y_i} + \frac{1}{2}\left(\ln\left(\phi\mu_i\right) - \ln\left(2\pi y_i^3\right)\right)\right) \tag{23}$$

No caso de a amostra ser truncada para valores das observações fora do intervalo $[c_1, c_2]$, a função de log-verosimilhança é

$$l = \sum_{i=1}^{n} \left(-\frac{\phi y_i}{2\mu_i} + \phi - \frac{\phi \mu_i}{2y_i} + \frac{1}{2}\left(\ln\left(\phi\mu_i\right) - \ln\left(2\pi y_i^3\right)\right) - \ln\left(G\left(c_2; \mu_i, \phi\right) - G\left(c_1; \mu_i, \phi\right)\right)\right). \tag{24}$$

Se a amostra for truncada para valores inferiores a c_1 e censurada para valores superiores a c_2, a função de log-verosimilhança será

$$\begin{aligned} l \;=\; & \sum_{i=1}^{n_0} \left[\ln\left(1 - G\left(c_2; \mu_i, \phi\right)\right) - \ln\left(1 - G\left(c_1; \mu_i, \phi\right)\right)\right] \\ & + \sum_{i=n_0+1}^{n_0+n_1} \left[-\frac{\phi y_i}{2\mu_i} + \phi - \frac{\phi \mu_i}{2y_i} + \frac{1}{2}\left(\ln\left(\phi\mu_i\right) - \ln\left(2\pi y_i^3\right)\right) - \ln\left(1 - G\left(c_1; \mu_i, \phi\right)\right)\right]. \end{aligned} \tag{25}$$

Tal como nas expressões das funções de log-verosimilhança que foram apresentadas para a distribuição gama, as expressões (23), (24) e (25) têm como hipótese que o valor esperado de cada observação é condicionado pelo valor que as variáveis explicativas assumem para essa observação, e que o parâmetro ϕ da distribuição IG2 é fixo para todas as observações.

Como valores iniciais dos parâmetros para o processo iteractivo de estimação de modelos com distribuição inversa gaussiana sugere-se, tal como para modelos gama, que se usem as estimativas do modelo lognormal equivalente. Os valores iniciais para modelos com distribuição inversa gaussiana e função de ligação logarítmica são, no caso da distribuição IG2, em tudo semelhantes aos do modelo gama, devido ao facto de os dois primeiros momentos das duas distribuições terem expressões idênticas.

Resultados da aplicação do modelo com distribuição inversa gaussiana a dados simulados À semelhança do que se fez para os modelos lognormal e gama, efectuaram-se diversas modelizações com dados simulados, para exemplificar a eficiência dos modelos tobit generalizados na estimação de parâmetros de distribuições do tipo inversa Gaussiana. Mais uma vez, geraram-se 10000 observações com distribuição inversa gaussiana, de parametrização IG2, com o valor esperado a ser condicionado pelas variáveis explicativas da seguinte forma:

$$\mu_i = \exp\left\{\beta_1 + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \beta_4 x_{i,4} + \beta_5 x_{i,5}\right\},$$

onde as variáveis explicativas têm as mesmas características que as utilizadas nas observações dos modelos lognormal e gama, e com

$$\beta_1 = -1,5 \quad \beta_2 = 0,5 \quad \beta_3 = 0,3 \quad \beta_4 = -0,3 \quad \beta_5 = -0,5 \ .$$

Na geração de todas as observações considerou-se que $\phi = 1,2$. Assim, assumiu-se um comportamento do valor esperado e da variância das observações igual ao utilizado para a distribuição gama.

Após se terem gerado as observações, foram estimados os parâmetros com diversas combinações de valores de truncagem e censura, cujos resultados estão patentes na Figura 7, onde a célula a sombreado sinaliza o parâmetro estimado em que se rejeita com um

	Truncagem	Censura					
		S/ Censura	1,00	0,60	0,40	0,20	0,15
β_1	S/ Truncagem	-1,491	-1,491	-1,490	-1,499	-1,499	-1,502
β_2		0,498	0,496	0,495	0,495	0,492	0,484
β_3		0,294	0,293	0,292	0,290	0,293	0,300
β_4		-0,308	-0,308	-0,309	-0,306	-0,300	-0,296
β_5		-0,503	-0,503	-0,499	-0,494	-0,492	-0,490
ϕ		1,190	1,193	1,192	1,205	1,195	1,191
β_1	0,03	-1,490	-1,490	-1,490	-1,497	-1,500	-1,504
β_2		0,494	0,491	0,490	0,488	0,484	0,475
β_3		0,306	0,305	0,303	0,301	0,306	0,316
β_4		-0,323	-0,323	-0,324	-0,320	-0,315	-0,312
β_5		-0,508	-0,507	-0,504	-0,497	-0,496	-0,493
ϕ		1,193	1,196	1,195	1,211	1,202	1,199
β_1	0,05	-1,497	-1,495	-1,494	-1,501	-1,503	-1,506
β_2		0,508	0,504	0,503	0,500	0,502	0,494
β_3		0,303	0,300	0,299	0,295	0,305	0,323
β_4		-0,313	-0,313	-0,315	-0,309	-0,301	-0,296
β_5		-0,514	-0,513	-0,509	-0,499	-0,502	-0,503
ϕ		1,178	1,183	1,181	1,201	1,172	1,150
β_1	0,10	-1,511	-1,503	-1,502	-1,504	-1,530	-1,577
β_2		0,539	0,530	-0,528	0,518	0,508	0,408
β_3		0,298	0,295	0,291	0,277	0,282	0,275
β_4		-0,324	-0,324	-0,328	-0,312	-0,280	-0,216
β_5		-0,516	-0,512	-0,503	-0,476	-0,455	-0,368
ϕ		1,189	1,202	1,205	1,270	1,325	1,749

Figura 7: Resultados de simulação com distribuição inversa Gaussiana

grau de confiança de 95% que seja igual ao seu verdadeiro valor.

Os resultados mostram que a estimação do modelo para observações de variáveis aleatórias com distribuição inversa Gaussiana é bastante precisa, apenas se verificando desvios significativos para truncagem em 0,1 e para as censuras em 0,2 e 0,15, situações em que o número de observações restringidas é assinalável. A quantidade de observações afectadas por cada uma das restrições impostas na estimação dos parâmetros pode ser consultada na Figura 8.

Mais uma vez, verifica-se que as estimativas dos parâmetros são mais sensíveis à truncagem da informação do que à censura, o que é compreensível, uma vez que a truncagem representa retirar toda a informação referente a algumas observações, enquanto que a censura representa que, para algumas observações, apenas se sabe que ocorreram e que o seu valor está dentro de determinado intervalo.

Truncagem					
Valor de Truncagem			0,03	0,05	0,10
Nº Observ. Truncadas			161	703	2589
Censura					
Valor de Censura	0,15	0,20	0,40	0,60	1,00
Nº Observ. Censuradas	5725	4463	1851	931	294

Figura 8: Nºde observações censuradas e truncadas - distribuição inversa Gaussiana

Os resultados indicam que a estimação dos modelos tobit generalizados com distribuição inversa Gaussiana é bastante menos problemática do que a estimação destes modelos com distribuição gama. No caso dos modelos inversa Gaussiana, as estimativas apresentam menores desvios, não se tendo verificado problemas de convergência no processo de estimação, e observando-se um erro de 1ªespécie em apenas uma das várias modelizações.

A qualidade dos resultados obtidos com a inversa Gaussiana deve-se em grande parte ao facto de no processo de estimação se ter utilizado uma função de distribuição que, conhecidos os valores da função de distribuição normal, é exacta, ao contrário do procedimento utilizado na estimação dos modelos com distribuição gama, onde se recorreu a uma função de distribuição aproximada, que também depende da função de distribuição da normal, e que se viu poder ser a origem de enviesamentos das estimativas dos parâmetros.

4 Limitações e pontencialidades das duas abordagens

A característica principal dos seguros consiste na transferência de um risco de um segurado para uma seguradora, a qual exige como contrapartida o pagamento de um prémio que traduz o preço do seguro.

Este risco poderá não gerar qualquer sinistro mas poderá também gerar um ou, em alguns tipos de seguros, mais sinistros. A cada sinistro gerado pelo risco está associada uma indemnização cujo valor é variável ou fixo, consoante o tipo de seguro. Assim, o valor total das indemnizações geradas por um risco i numa anuidade, designado por processo de risco, pode ser entendido como uma variável aleatória Z_i que tem origem num processo composto

$$Z_i = \sum_{j=0}^{N_i} Y_{ij},$$

onde N_i representa o número de sinistros participados numa anuidade e Y_{ij} indica o valor da indemnização associada à j-ésima participação, com $Y_0 \equiv 0$. Assumindo que, para cada risco i, as indemnizações Y_{ij} são independentes de N_i e constituem uma sucessão de variáveis aleatórias independentes e identicamente distribuídas, então

$$E[Z_i] = E[N_i] \times E[Y_i], \tag{26}$$

onde Y_i é uma variável aleatória com a distribuição comum a todos os Y_{ij}. Nestas condições, o valor esperado de Z_i corresponde ao número esperado de sinistros multiplicado pelo valor esperado de cada indemnização individual.

Como o prémio é pago no início da anuidade e o valor das indemnizações dessa anuidade, Z_i, só é conhecido posteriormente, o valor do prémio deve ser definido com base na distribuição das indemnizações agregadas de uma anuidade de determinado risco.

Uma análise das diversas funções do prémio de seguro mostra que este pode ser desagregado em duas componentes com diferentes finalidades. A parcela principal do prémio destina-se a fazer face às indemnizações geradas pelo risco. A outra parcela relaciona-se com custos associados ao exercício da actividade seguradora, nomeadamente, os custos de gestão dos contratos e dos sinistros, os custos de aquisição dos contratos, a carga fiscal e a remuneração do capital investido na actividade. Os valores que estas duas parcelas devem assumir, para determinado risco, têm que ser estimados recorrendo a abordagens diferenciadas. Neste trabalho, apenas se abordam os aspectos associados com o cálculo da primeira parcela dos prémios, ou seja, da componente de risco.

Em (26) pressupõe-se que se conhece o valor esperado do número de sinistros e de cada indemnização, para determinado risco i. Se bem que para alguns riscos, de alguns tipos de seguros, se possam ter estimativas de $E[N_i]$ e $E[Y_i]$, regra geral, não existe informação suficiente para formar uma expectativa sobre a sinistralidade específica de cada risco individual.

Se se considerar uma carteira de apólices composta por um conjunto de riscos que, apesar de heterogéneos, apresentam um comportamento com algumas semelhanças e onde cada risco é caracterizado por um parâmetro θ_i (uni ou multi dimensional), pode-se assumir que as distribuições de N_i e Y_i, para todos os riscos, pertencem a uma mesma família de distribuições, sendo apenas diferenciadas pelos parâmetros θ_i associados a cada risco. Nestas condições,

$$E[Z|\theta_i] = E[N|\theta_i] \times E[Y|\theta_i],$$

onde Z, N e Y são variáveis aleatórias. Assim, está-se a assumir que as indemnizações de riscos que partilham o mesmo parâmetro θ têm o mesmo comportamento esperado de frequência e de custo.

Em termos práticos, o parâmetro θ varia em função dos valores assumidos pelos factores tarifários referentes a cada risco.

Os modelos apresentados nas dois capítulos anteriores são ferramentas importantes na estimação dos prémios para riscos pertencentes a carteiras com um elevado número de apólices e um comportamento não muito heterogéneo, porque podem, nomeadamente, ser aplicados quer a distribuições de contagem, quer a distribuições continuas. Assim, podem ser utilizados para modelizar o número de sinistros ocorridos e o valor das indemnizações de cada sinistro.

Muito embora os modelos lineares generalizados se mostrem desadequados para tratar os custos dos sinistros, quando eles envolvem franquias e limites de indemnização, são utilizados de forma genérica em diversos tipos de seguros com muitas unidades em risco e com um reduzido grau de heterogeneidade. Neste capítulo expõem-se as razões dessa desadequação e mostra-se como os modelos tobit generalizados podem ultrapassar algumas insuficiências dos modelos lineares generalizados.

4.1 Franquias e limites de indemnização

A franquia e o limite de indemnização são duas característas exógenas das apólices de seguro que têm bastante influência sobre o valor agregado das indemnizações geradas por um risco, numa anuidade.

4.1.1 Franquias

A franquia é definida como a parcela dos prejuizos indemnizáveis gerados por um sinistro que fica a cargo do segurado. As franquias podem assumir diversas formas, e têm impactos diferentes no número de sinistros participados à seguradora e nos montantes destes sinistros.

Em quase todos os tipos de seguros podem-se considerar franquias. No entanto, em

alguns casos, dadas as características do seguro, é pouco usual a sua aplicação. Os seguros de Responsabilidade Civil obrigatórios, nomeadamente o de Automóvel, são exemplos de seguros em que é pouco frequente aplicarem-se franquias, principalmente porque são seguros em que se paga a indemnização a um terceiro, ao qual não se pode cobrar uma franquia.

Apesar de, em alguns tipos de seguros, as franquias serem iguais para todos os contratos, existem seguros em que o valor da franquia depende da opção do segurado, podendo este escolher uma franquia mais elevada, ou mais reduzida, por forma a que esta se ajuste ao seu perfil financeiro e de aversão ao risco. Em alguns seguros pode também optar-se pela não aplicação de franquias. Note-se que, quando o segurado opta por um franquia superior tem como contrapartida uma redução no prémio a pagar.

A heterogeneidade das franquias dos diversos contratos pode conduzir a que a seguradora observe, para riscos com características semelhantes, comportamentos de sinistralidade diferenciados, mesmo considerando várias anuidades.

Refira-se que, em seguros em que estejam garantidas diversas coberturas, a franquia aplicável pode variar com a cobertura, como acontece, por exemplo, nos seguros de Riscos Multiplos Habitação onde é frequente encontrarem-se, no mesmo contrato, algumas coberturas sem franquia, e outras coberturas com franquias, assumindo estas diversas formas e montantes.

Actualmente verifica-se uma tendência crescente para se eliminarem as franquias em contratos que tradicionalmente previam a sua aplicação. Por exemplo, são cada vez mais as seguradoras que apresentam coberturas de Danos Próprios com a opção de não aplicação de franquias, enquanto até há alguns anos atrás essa alternativa não estava prevista. Esta tendência deve-se principalmente a três factores: a crescente apetência dos segurados por contratos sem franquia (apesar do incremento no preço), a vontade das seguradoras

em melhorarem as relações com os clientes e uma maior simplicidade no processo de regularização dos sinistros.

Como os diversos tipos de franquias têm diferentes impactos sobre o comportamento da sinistralidade e consequentemente sobre o cálculo do prémio, é de seguida apresentada uma classificação das franquias relativamente a aspectos que se considera serem os mais relevantes para este trabalho:

1. Quanto à dedutibilidade:

- Dedutível: quando a indemnização corresponde ao prejuizo indemnizável subtraído do valor da franquia. Neste tipo de franquia uma parte do prejuizo indemnizável é sempre suportado pelo segurado.

- Não dedutível: o valor da franquia corresponde apenas ao prejuízo mínimo indemnizável; quando o prejuízo é inferior à franquia não há lugar a indemnização por parte da seguradora, enquanto que, se exceder a franquia, a indemnização é paga integralmente.

2. Quanto à variabilidade:

- Variável em função do sinistro: o valor da franquia depende de características do sinistro, geralmente, do valor do prejuízo indemnizável. Por exemplo, 5% do prejuizo indemnizável. O seu valor só é conhecido no momento do encerramento do sinistro. Quando é puramente uma percentagem do prejuízo indemnizável, pode ser entendida como um co-seguro com o segurado.

- Variável em função das características do contrato: varia de contrato para contrato em função de características do contrato, sendo geralmente uma percentagem do limite de indemnização. Por exemplo, 2% do limite de indemnização. Neste tipo

de franquias o segurado sabe que terá de suportar os sinistros que gerem prejuizos inferiores à franquia pré-determinada.

- Fixa: tal como no caso anterior, o seu valor é conhecido no momento da celebração do contrato de seguro mas é definido de forma completamente independente das características do contrato. Por exemplo, 500 euros.

3. Quanto à complexidade:

- Simples: o valor da franquia é determinado exclusivamente por uma condição, como por exemplo, uma franquia dedutível em percentagem do prejuízo indemnizável ou uma franquia de valor fixo.

- Complexa: quando o valor da franquia é o resultado de duas ou mais condições, por exemplo: uma franquia de 2% do limite de indemnização com um mínimo 100 euros e um máximo de 500 euros.

Algumas das combinações de características das franquias não fazem sentido, como por exemplo, uma franquia em percentagem do prejuízo indemnizável, não dedutível.

Como facilmente se entende, os impactos das franquias sobre o valor das indemnizações e sobre o número sinistros participados podem ser diversos. Uma franquia simples, dedutivel e em percentagem do prejuízo indemnizável, não tem qualquer impacto sobre o número de sinistros participados, uma vez que qualquer que seja o valor do prejuízo, há sempre lugar a uma indemnização, enquanto o valor das indemnizações virá sempre inferior ao prejuízo indemnizável. No caso de se ter uma franquia não dedutível, fixa e simples, os seus efeitos são, de certa forma, opostos aos do caso anterior: o número de sinistros declarados poderá ser inferior ao número de ocorridos, uma vez que não são indemnizáveis prejuízos de valor inferior à franquia, e, para os sinistros declarados, o valor da indemnização corresponde ao prejuízo indemnizável.

Por ser o caso de maior interesse, daqui em diante, tratar-se-á o caso da franquia dedutível, fixa e simples, excepto quando se refira outra situação.

4.1.2 Limites de indemnização

Os limites de indemnização são uma das características mais importantes das apólices de seguro, uma vez que definem os limites de responsabilidade que são assumidos pela seguradora em caso de sinistro. Poucos são os tipos de seguro em que a responsabilidade da seguradora não está limitada, sendo a Responsabilidade Civil Automóvel um desses poucos casos. Em quase todos os restantes tipos de seguro existe um limite de indemnização que define a responsabilidade da seguradora.

Ao analisar os principais tipos de seguros encontram-se três grupos distintos: os seguros que cobrem acidentes e doenças em pessoas, os seguros que cobrem a responsabilidade civil e os seguros que cobrem os danos patrimóniais. Nos seguros de acidentes e doença, os limites de indemnização podem, dependendo do tipo de seguro, assumir diferentes formas. Salvo algumas excepções, não se consegue uma correcta modelização dos custos desses tipos de seguros através dos modelos em estudo, modelos lineares generalizados e tobit generazados, pelo que se prestará apenas atenção aos outros dois grupos de seguros, por terem maior interesse para este trabalho.

Nos seguros de danos patrimoniais e de responsabilidade civil a natureza do limite de indemnização apresenta diferenças substanciais. Enquanto que nos seguros sobre os danos patrimoniais o limite de indemnização tem uma correspondência fisica, ou seja, o valor do bem seguro, no caso dos seguros de responsabilidade civil, o limite de indemnização é subjectivo, sendo geralmente definido por forma a que grande parte das indemnizações sejam abrangidas.

Nos seguros de danos patrimoniais o procedimento usual é a reparação dos bens afecta-

dos. Quando o custo de reparação excede o valor seguro, ou o bem seguro não é reparável, indemniza-se o segurado pelo valor do bem afectado. Neste tipo de sinistros, que se classificam de perda total, atinge-se o limite de indemnização da apólice, situação que ocorre com alguma frequência. Em seguros como o de Incêndio, a probabilidade de ocorrência de uma perda total, dado que ocorreu um sinistro, pode ser considerável.

No caso dos seguros de Responsabilidade Civil, como o limite de indemnização é definido subjectivamente, a probabilidade de um sinistro atingir o limite da apólice depende muito do tipo de risco e do limite acordado. Regra geral, no caso dos seguros de responsabilidade civil obrigatórios, o capital mínimo exigido por lei é definido por forma a que probabilidade de uma indemnização exceder o limite seja mínima. Um bom exemplo desta situação é, novamente, a Responsabilidade Civil Automóvel onde o capital mínimo destas apólices é 120 mil contos e ao longo de vários anos foram muito poucos os casos que excederam esse valor. Quando se consideram outros seguros de responsabilidade civil, como por exemplo Responsabilidade Civil Exploração, onde se subscrevem apólices com limites de indemnização de 5 mil contos, já é mais provável a ocorrência de um sinistro que atinja o capital seguro.

Em termos práticos, os seguros podem ser classificados quanto ao limite de indemnização como, seguros em que existe uma probabilidade não desprezível de que um sinistro gere um prejuízo indemnizável de valor igual ou superior ao limite e seguros em que o limite de indemnização é ilimitado (não existe limite) ou é improvável de ser atingido, por ser muito pequena a probabilidade de uma indemnização exceder o limite previsto na apólice.

4.2 Insuficiências dos modelos lineares generalizados

Os modelos lineares generalizados assistiram, ao longo da última década, a uma crescente utilização na estimação de estruturas tarifárias. A utilização destes modelos contribuiu de forma significativa para uma maior adequação dos prémios de diferentes segmentos de risco em carteiras com um elevado número de apólices. Em grande parte esse sucesso deveu-se:

- às vantagens de ser uma análise multivariada dos factores de tarifação;
- ao facto de algumas distribuições da família de dispersão exponencial se ajustarem bem quer ao número de participações (como a distribuição de Poisson), quer aos custos (como a distribuição gama, por exemplo) de determinados tipos de seguros;
- por serem modelos relativamente simples, com um tratamento integrado semelhante para diversas distribuições;
- por o método de estimação ser rápido, permitindo a estimação de modelos com um grande número de observações e muitos factores de tarifação, o que levou ao desenvolvimento de um mercado de *software* específico para tarifação, baseado nestes modelos;
- à boa adaptação das características dos modelos ao seguro de Responsabilidade Civil Automóvel, o seguro de maior importância na generalidade dos mercados, e ao qual são aplicados com maior frequência.

Os modelos lineares generalizados adaptam-se bem ao seguro de Responsabilidade Civil Automóvel por este ser um tipo de seguro:

- com um grande número de unidades em risco;
- composto por riscos com um grau de heterogeneidade relativamente baixo;

- em que o número de sinistros de cada segmento de riscos é bem ajustado por uma distribuição de Poisson;
- para o qual os custos são relativamente bem modelizados por distribuições gama ou lognormal.

A modelização dos custos através de uma dessas distribuições fornece, regra geral, bons resultados. Um dos factores que contribui para a qualidade do ajustamento é o facto de se tratarem de distribuições com domínio em $\Re^+$ que pressupõem que não existe limite para o valor da indemnização, o que é uma hipótese aceitável uma vez que, na prática, se verifica (por ser improvável que o limite de indemnização seja alcançado).

Outro factor que contribui para a adequação dos modelos lineares generalizados, na modelização do valor da indemnização, é o reduzido número de contratos com franquia nos seguros de Responsabilidade Civil Automóvel. Apesar de ser possível que este tipo de contratos preveja a aplicação de uma franquia, é pouco frequente que se subscrevam contratos nessas condições, principalmente por a franquia não ser oponível aos lesados.

Um exemplo de aplicação dos modelos lineares generalizados para a construção de uma tarifa de responsabilidade civil automóvel pode ser encontrado em Barata (2000).

Recentemente, tem-se assistido à utilização dos modelos lineares generalizados para a estimação de tarifas de outros tipos de seguros. Esses seguros têm, geralmente, características diferentes face aos seguros de Responsabilidade Civil Automóvel, nomeadamente, é usual terem franquias e apresentarem uma probabilidade assinalável de que o limite máximo de indemnização seja atingido. Estas diferenças nas características dos seguros, conduzem a que os modelos lineares generalizados sejam menos adequados para modelizar o comportamento das indemnizações, do que acontece nos seguros de responsabilidade civil automóvel. Os seguros de Danos Próprios Automóvel e de Riscos Multiplos Habitação são os principais tipos de seguros nessas condições.

O procedimento usual para modelizar os custos na presença de franquias consiste, basicamente, em considerar como variável endógena o valor da indemnização suportado pela seguradora (indemnização ou custo de reparação eventualmente deduzido da franquia) e incluir como variável de tarifação o valor da franquia. Este procedimento dá resultados razoáveis quando se pretende estimar o impacto de franquias já existentes e, no futuro, apenas se pretende manter esses valores de franquia. No caso de se ter como objectivo alterar o tipo de franquia, ou apenas alterar os valores de franquia disponíveis (por exemplo, passar a disponibilizar uma opção de seguro sem franquia), esta modelização fornece informação insuficiente.

A existência de limites máximos de indemnização não é objecto de qualquer tratamento específico na modelização do valor das indemnizações através dos modelos lineares generalizados. Este facto leva a que, quando a probabilidade de um sinistro provocar a indemnização máxima é significativa, a estimativa do valor esperado das indemnizações tenha um enviesamente positivo. Esse enviesamento é crescente com a probabilidade de perda total.

Por estas razões se defende que os modelos lineares generalizados não são adequados para modelizar indemnizações com as características atrás descritas. Os modelos tobit generalizados, por permitirem a modelização de variáveis com base em amostras truncadas e censuradas, são mais ajustados a essas situações, sendo utilizados, neste trabalho, para ultrapassar as limitações dos modelos lineares generalizados.

4.3 A modelização dos custos com os modelos tobit generalizados

Considere-se que determinada apólice tem uma franquia dedutível e fixa de valor D e que o seu limite de indemnização é L. Se o valor de cada indemnização suportada pela seguradora, para essa apólice, for X, então X assumirá valores no intervalo $[0; L - D]$.

Seja Y a variável aleatória que representa o valor do prejuizo que deu origem a X. Como, geralmente, a seguradora regista apenas o valor das indemnizações pagas (observações de X), ela observará valores de Y no intervalo $[D; L]$ porque os segurados não participarão sinistros com prejuízo inferior à franquia e sempre que o prejuízo for superior ao limite de indemnização, a seguradora só indemniza até ao valor limite considerado na apólice. Isto significa que a amostra de custos resultantes de sinistros observada pela seguradora apresenta dois tipos de restrições: devido à existência de franquias, é truncada à esquerda, uma vez que a seguradora não toma conhecimento dos sinistros de valor inferior à franquia; devido à existência do limite de indemnização, algumas observações são censuradas à direita.

Assim, o tipo de amostra que resulta de uma carteira de apólices com as características enunciadas enquadra-se no âmbito dos modelos tobit generalizados, os quais permitem estimar os parâmetros da distribuição Y subjacente às indemnizações, em função dos factores de tarifação. Os valores de Y corresponderiam às indemnizações caso não existissem nem franquias, nem limites de indemnização.

Uma vez estimados os parâmetros da função de distribuição de Y, para um risco com determinados factores de tarifação, está-se na posse de informação suficiente para se calcular o valor esperado das indemnizações, dada a existência de uma qualquer combinação de franquia e limite de indemnização.

Em termos formalizados, o valor da indemnização X, quando existe uma franquia dedutível D, depende do valor do custo subjacente, Y, da seguinte forma

$$X = \begin{cases} 0 & Y < D \\ Y - D & D \leqslant Y < L \\ L - D & Y \geqslant L \end{cases} .$$

Como o valor que a seguradora normalmente regista, para cada sinistro, é X, basta somar D, que é um valor conhecido, para se reconstituir o valor de Y (após truncagem e censura).

Se se considerar uma a franquia não dedutível, pode-se de igual forma utilizar os modelos tobit generalizados para estimar a distribuição subjacente, sendo apenas necessário ter em atenção que o valor da indemnização paga, X^*, é agora expresso em termos de Y como

$$X^* = \begin{cases} 0 & Y < D \\ Y & D \leqslant Y < L \\ L & Y \geqslant L \end{cases} .$$

No Capítulo 3 apresentaram-se, para as várias distribuições consideradas, as expressões de $E[\min(Y, c_2)|Y > c_1]$. Considerando que $c_1 = D$ e que $c_2 = L$, então

$$E[\min(Y, c_2)|Y > c_1] = E[X^*|Y \geqslant D] .$$

Facilmente se verifica que

$$E[X^*] = E[X^*|Y \geqslant D]\,[1 - \Pr(Y \leqslant D)] . \tag{27}$$

Outras relações importantes para o cálculo das tarifas, referentes a valores esperados de X e X^*, que facilmente podem ser expressos em termos de $E[X^*]$ são

$$E[X] = E[X^*] - D\,[1 - \Pr(Y \leqslant D)] \tag{28}$$

e

$$E[X|Y \geqslant D] = E[X^*|Y \geqslant D] - D = \frac{E[X^*]}{1 - \Pr[Y \leqslant D]} - D. \tag{29}$$

As expressões apresentadas nesta secção correspondem a franquias simples com valor fixo. Este caso pode facilmente ser generalizado para franquias em proporção do limite de indemnização. Quando se está na presença de outros tipos de franquias, nomeadamente franquias complexas, as expressões têm de ser adaptadas a cada situação específica.

5 Estimação da distribuição de custos com sinistros

Nos capítulos anteriores apresentaram-se alguns métodos de estimação paramétricos e referiu-se a utilidade que têm no cálculo de tarifas, nomeadamente na modelização dos custos com sinistros. Tal como se referiu no Capítulo 4, os modelos lineares generalizados adaptam-se bem à modelização dos custos de alguns tipos de seguros, entre os quais se destaca o seguro de Responsabilidade Civil Automóvel. No entanto, existem outros tipos de seguros cujos custos com sinistros têm características que não são bem modelizáveis através dos modelos lineares generalizados, sendo proveitoso o recurso às metodologias baseadas nos modelos tobit generalizados, que se apresentaram no Capítulo 3.

Neste capítulo, ilustra-se a utilização dos modelos tobit generalizados para modelizar a distribuição de custos com sinistros de um tipo de seguros que, pelas suas características, não é adequadamente modelizado pelos modelos lineares generalizados, embora estes sejam, frequentemente, a "ferramenta" escolhida para esse fim.

Para esta ilustração, escolheu-se a cobertura de Choque, Colisão ou Capotamento do seguro Automóvel. Esta deverá ser a segunda cobertura em que mais se utilizam os modelos lineares generalizados, logo a seguir à cobertura de Responsabilidade Civil.

5.1 Características da cobertura de Choque, Colisão ou Capotamento

A cobertura de Choque, Colisão ou Capotamento (CCC), é uma das coberturas facultativas do seguro automóvel. A cobertura de CCC é usualmente contratada simultaneamente com as coberturas de Furto ou Roubo e de Incêndio, Raio ou Explosão, e em conjunto formam o que tradicionalmente se designa por seguro de Danos Próprios[1]. Por política de subscrição das seguradoras, estas coberturas só podem ser contratadas quando também se subscreve

[1] Apesar de estas três coberturas serem tradicionalmente subscritas em conjunto, é usual poder-se subscrever apenas algumas delas. Note-se que é vulgar estarem também disponíveis outras coberturas de Danos Próprios.

o seguro obrigatório de responsabilidade civil automóvel. Apesar da cobertura de CCC garantir danos com origem em três tipos de acontecimentos diferentes, o prémio de seguro nunca é definido para cada um desses acontecimentos, mas sim para toda a cobertura. Podem-se definir os eventos cobertos da seguinte forma:

- Choque: danos no veículo seguro resultantes do embate contra qualquer corpo fixo, ou sofridos por aquele quando imobilizado;
- Colisão: danos no veículo seguro resultantes do seu embate com qualquer outro corpo em movimento;
- Capotamento: danos no veículo seguro decorrentes da perda da sua posição normal mas não decorrentes de choque ou colisão.

O cálculo do prémio *a priori* da cobertura de CCC pode ser efectuado de diversas formas. No entanto, é prática corrente, considerar como o factor mais importante na definição do prémio, o valor do veículo. Usualmente, outros factores são também tidos em conta, os quais variam de seguradora para seguradora, destacando-se a idade do condutor, a idade da carta, as características do veículo (que podem incluir a cilindrada, a potência, o peso e a idade) e a zona de circulação.

Apesar de se verificarem algumas excepções, o prémio é calculado como a multiplicação do valor do veículo por uma taxa constante para todos os veículos, a qual está depois sujeita a descontos e agravamentos que dependem das outras características do veículo e do condutor considerados. Esta fórmula de cálculo do prémio implica que, mantendo tudo o resto constante, o prémio é proporcional ao valor seguro.

Regra geral, o prémio da cobertura de CCC está também sujeito a tarifação *a posteriori*, ou seja, depende da experiência de sinistralidade do segurado, através de um sistema de bonus-malus, que, quando existe, é o mesmo que afecta a cobertura de responsabilidade

civil. A utilização de um sistema de bonus-malus, tem como consequência prática um comportamento que se designa por *sede de bonus* e que consiste na não participação de sinistros que gerem uma pequena indemnização, para evitar que o prémio seja agravado.

Na cobertura de CCC, o limite de indemnização corresponde ao valor seguro, o qual deve ser o valor venal do veículo à data da contratação da apólice, sendo a sua definição da competência do segurado. No caso de o valor seguro, declarado pelo segurado, ser inferior ao real valor venal do veículo, em caso de acidente, a seguradora paga a indemnização pela parte proporcional dos danos, correspondente à percentagem do capital seguro em relação ao valor venal. Quando o valor seguro é superior ao valor venal a seguradora apenas indemniza até ao valor venal.

Neste tipo de seguros é usual que após a ocorrência de um sinistro, o montante da indemnização seja abatido ao valor seguro, ficando este reduzido daquele valor desde a data do sinistro até ao vencimento anual do contrato. No entanto, o segurado pode repor o capital através do pagamento de um prémio suplementar correspondente ao capital reposto e ao período de tempo não decorrido, até ao vencimento do contrato.

Apesar de existirem alguns casos em que os contratos não têm franquia, a generalidade dos contratos e dos produtos com esta cobertura prevêm a aplicação de uma franquia em caso de indemnização por sinistro. Os produtos comercializados no mercado nacional disponibilizam diversas hipóteses de franquia, ficando ao critério do segurado a escolha de um valor adequado. Ao optar por um valor de franquia superior à franquia *standard*, o segurado beneficia, naturalmente, de uma redução no prémio.

Em todos os casos conhecidos no mercado português, a franquia é dedutível. Geralmente a franquia é variável em função do valor seguro, existindo também alguns casos em que ela é fixa e outros, embora menos frequentes, em que é variável em função do valor do prejuízo indemnizável. Normalmente as franquias são simples, não possuindo limites

máximos nem mínimos. Quando a franquia é variável em função do valor seguro, ela é definida como uma percentagem desse valor, sendo 2% o valor mais frequente.

O custo da seguradora com um sinistro corresponde ao valor do prejuízo indemnizável (cujo valor máximo depende do valor venal) adicionado dos custos de peritagem (remuneração do perito que avalia os danos no veículo e define um valor "razoável" para a reparação do veículo) e subtraído da franquia e, eventualmente, do valor do salvado.

O Decreto-Lei nº2/98, que procedeu à revisão de diversas disposições do Código da Estrada, define como salvado o veículo que "tenha sofrido danos que afectem gravemente as suas condições de segurança" ou "cujo valor de reparação seja superior a 70% do valor venal do veículo à data do sinistro". Apesar do que é definido neste Decreto-Lei, a seguradora e o segurado podem acordar que um veículo seja reparado mesmo que o seu custo represente mais de 70% do valor venal.

Caso o segurado fique com o salvado, o seu valor é deduzido ao valor da indemnização que lhe é paga. Se é a seguradora que fica na posse do salvado, procede à sua venda, ficando o custo final da seguradora igual ao valor da indemnização deduzido da receita resultante da venda do salvado.

Quando não se procede à reparação do veículo acontece aquilo que se designa por *perda total*. Nos casos em que é económica e tecnicamente viável a reparação do veículo, ocorre uma *perda parcial*.

5.2 A informação utilizada

5.2.1 Características dos sinistros

Para a aplicação dos modelos tobit generalizados na estimação da distribuição de custos com sinistros utilizou-se, como já se mencionou, informação referente à cobertura de Choque, Colisão ou Capotamento do seguro automóvel da carteira de uma seguradora em

actividade no mercado português.

A definição do período abrangido pela informação analisada é um aspecto de grande importância. Por um lado é conveniente ter um elevado número de sinistros para que a amostra seja representativa, mas, simultaneamente, não convém que sejam considerados sinistros muito antigos, pois as características dos veículos e o valor das reparações variam no tempo. No caso específico da cobertura de CCC, é necessário ter presente que em 1998 se deu inicio à aplicação do Decreto-Lei n°214/97 de 16 de Agosto, que determina a actualização anual do valor do veículo seguro (excepto se, por acordo entre o segurado e a seguradora, se definir outro valor seguro). Dado que os critérios de valorização dos veículos foram alterados em 1998, não é adequado utilizar-se simultanamente informação anterior e posterior a esse ano.

Em função destas restrições, optou-se por utilizar sinistros ocorridos e declarados no período decorrido entre 01-01-1998 e 31-08-2000. A informação recolhida corresponde à situação dos respectivos processos a 31-08-2000, o que significa que alguns sinistros ainda não se encontravam encerrados nesse momento (aproximadamente 20% dos sinistros).

Como os sinistros não encerrados têm uma provisão casuística que reflecte uma estimativa do custo que a seguradora ainda terá de suportar com o sinistro, baseada na avaliação dos danos por um perito, considerou-se que o custo final do sinistro corresponde aos montantes já pagos mais a provisão para indemnizações. Para o processo de estimação, foram retirados os sinistros encerrados com custo nulo e aqueles que, não estando ainda encerrados, se espera que não venham a gerar custos.

Apesar de a carteira ter apólices com todo o tipo de veículos, só foram considerados sinistros referentes a apólices segurando veículos ligeiros (peso bruto inferior a 3500 Kg e número de passageiros inferior ou igual a 9).

Para uma correcta caracterização da unidade de risco e dos factores que influenciam

o comportamento dos custos, foi recolhida informação da base de dados dos contratos, que não estava incluída na base de dados dos sinistros. Como não foi possível recolher a informação dos contratos à data exacta da ocorrência do sinistro, recorreu-se à informação dos contratos em 31 de Dezembro de cada ano. No caso dos sinistros de apólices com início no ano da ocorrência, recolheu-se informação referente ao fim desse ano, enquanto que nos restantos casos, optou-se pela informação do estado da apólice a 31 de Dezembro do ano anterior.

Por se utilizar a informação, associada aos contratos, registada em momentos diferentes daqueles em que ocorreram os sinistros, podem verificar-se desvios entre as características do veículo sinistrado e aquelas que são consideradas no processo de estimação. Apesar de esta situação perturbar a modelização dos custos com sinistros, admite-se que, quando ocorrem alterações de veículo, o novo veículo tem características próximas do antigo; nestas condições, os erros na informação têm impactos limitados.

A análise da informação permitiu constatar que existem sinistros cujo custo excede claramente o valor do veículo seguro, registado na base de dados de apólices. Por se suspeitar que essas situações são resultado de problemas na informação, foram excluídos da análise todos os sinistros cujo custo seja superior a 130% do valor seguro, num total de 58 sinistros, com o objectivo de evitar que tenham impactos nos parâmetros estimados. Consideraram-se sinistros até 130% do valor seguro, e não apenas 100%, por causa da existência de custos de peritagem e por ser possível o segurado e a seguradora acordarem a reparação de um veículo muito danificado, situações que podem implicar que o custo de um sinistro exceda o valor seguro.

Após se terem retirado os sinistros com custo nulo e aqueles cujo custo representa uma proporção excessiva do valor seguro, ficaram os 21071 sinistros que se vão utilizar nas diversas modelizações.

Na carteira de apólices de onde foram extraídos os sinistros em estudo, apenas um pequeno número de contratos não prevêm a dedução de uma franquia, em caso de sinistro. Os restantes contratos prevê a aplicação de uma franquia simples e em percentagem do valor seguro; essa percentagem é variável de contrato para contrato. Verificou-se que, todos sinistros utilizados neste trabalho foram gerados por apólices com franquia. Como o valor da franquia deduzida depende do valor seguro e da percentagem aplicável, este variou de sinistro para sinistro. Na grande maioria dos casos a franquia representou 2% do capital seguro.

5.2.2 Características da base de dados

A base de dados, de onde foi retirada a informação, possui diversas características que influenciam a qualidade dos resultados finais e que obrigam a que se tenham alguns cuidados na sua análise.

A informação foi extraída de uma base de dados antiga, que foi construida quando a capacidade de armazenamento de informação era bastante reduzida. Para além disso, nessa altura, não se faziam sentir grandes necessidades de registar muita informação, não só por não se considerar a informação como uma vantagem competitiva no negócio, mas também por não existirem ferramentas que permitissem analisar os dados de forma expedita, ao contrário do que sucede hoje. Como consequência, o volume de informação disponível para cada contrato é reduzido, estando limitado à informação mais relevante para a gestão do contrato.

Uma outra característica da base de dados refere-se à qualidade da informação registada. Nem sempre houve uma grande preocupação com a qualidade dos dados recolhidos, principalmente se se referiam a variáveis sem implicação directa na gestão ou tarifação dos contratos, como sucede, por exemplo, com o sexo do condutor. Esta situação tem como

consequência que, em alguns campos, haja informação omissa ou estejam introduzidos valores por defeito e, noutros casos, a informação não seja muito precisa.

Apesar de se sentir que este cenário se está a alterar, as características descritas são partilhadas pelas bases de dados de muitas seguradoras portuguesas.

A pouca quantidade e qualidade de informação condiciona os resultados. O facto de a quantidade de informação disponível sobre os diversos riscos ser reduzida, impossibilita a utilização, na modelização, de algumas variáveis que se pensa influenciarem o comportamento dos custos. Como essas variáveis não estão incluidas, o seu efeito será associado a outras, principalmente àquelas que no conjunto dos contratos têm maior correlação com as variáveis omissas.

Tornar a base de dados mais completa e com informação de maior qualidade careceria de contactos com os clientes, para que validassem a informação existente e fornecessem outra suplementar, o que envolveria muitos recursos financeiros e humanos. Como o objectivo deste trabalho não é construir uma tarifa para a cobertura de CCC, mas sim ilustrar, com dados reais, como os modelos tobit generalizados podem ser uma "ferramenta" importante na modelização dos custos e na determinação de impactos sobre a frequência de sinistralidade, a informação disponível considera-se suficiente.

5.2.3 Variáveis analisadas

Dadas as características da informação disponível, foram utilizadas algumas variáveis para modelizar o comportamento dos custos, as quais se passam a descrever.

"Sexo" do condutor habitual: Esta variável apresenta três valores possíveis: Masculino, Feminino e Empresa. Esta variável não tem impacto sobre a gestão ou tarifação dos contratos, o que leva a que se suspeite da existência de alguns erros no registo da informação, nomeadamente que nalguns contratos esteja incorrectamente introduzido o

valor por defeito, que corresponde a Empresa.

Categoria do veículo: Foram utilizadas, na análise, todas as categorias de veículos ligeiros com representatividade na carteira. A utilização das diferentes categorias de veículos foi condicionada pela classificação utilizada no registo da informação. Assim, utilizaram-se as seguintes categorias: ligeiros de passageiros, ligeiros mistos, caminhetas, veículos de todo-o-terreno e veículos em regime de leasing ou ALD.

Esta última categoria é composta por veículos das restantes categorias, pois todos os veículos adquiridos em regime de leasing ou ALD foram registados nessa categoria. Assim, se esta categoria tiver um impacto significativo sobre o comportamento dos custos com sinistros, esse impacto terá de ser interpretado como estando associado às características dos clientes que recorrem a estes regimes de aquisição e ao tipo de veículos que são adquiridos dessa forma.

Peso Bruto: A informação referente ao peso bruto dos veículos não está registada, na base de dados, em valor, mas sim por intervalos de peso. As classes utilizadas no registo da informação foram: até 1600 Kg, de 1601 a 2900 Kg e de 2901 a 3500Kg. Como se encontrou uma proporção elevada de veículos sem registo do peso bruto, tornou-se necessário acrescentar a estas três classes, uma quarta classe onde se incluiram os veículos para os quais se desconhece o peso bruto.

Esta variável apresenta uma forte correlação com a categoria do veículo, uma vez que uma das características que define a categoria do veículo é o seu peso bruto.

Cilindrada do Veículo: Tal como o peso bruto do veículo, esta variável está registada, na base de dados, por classes, sendo elas as seguintes: até 1500 cc, de 1501 a 2500 cc e mais de 2500 cc. Estas foram as três classes utilizadas na modelização dos custos.

Idade do Veículo: Esta variável foi construída através do ano de construção do veículo, o qual se encontra registado na base de dados. Esta informação vai ser considerada como variável qualitativa. Os veículos foram classificados em onze classes: uma por cada ano entre zero e nove anos e uma outra classe para veículos com mais de nove anos.

Valor Seguro: O capital seguro deve coincidir com o valor venal do veículo. Salvo situações em que, por acordo entre o segurado e a seguradora, se defina outro capital, o valor seguro corresponde a uma proporção do valor do veículo em novo, a qual representa o impacto da desvalorização do veículo e se encontra tabelada. Ao contrário de todas as outras variáveis esta não é uma variável qualitativa, mas sim quantitativa. Refira-se que, apesar de ser uma variável quantitativa, é prática comum criar-se um elevado número de classes com base nos valores seguros e utilizar esta variável como qualitativa.

Franquia: Entre as características do segurado que influenciam a escolha do valor da franquia, deverão estar o seu perfil de aversão ao risco, a sua capacidade financeira e também as suas expectativas quanto ao seu próprio comportamento de sinistralidade. Por forma a tentar captar eventuais diferenças do comportamento dos custos com sinistros de segurados que optam pelas diferentes franquias, introduziu-se essa variável na estimação. Foram criadas 5 classes de franquia, correspondentes aos diversos tipos de franquia existentes: 2%, 4%, 8%, 12% e 20% do valor seguro.

Idade do Condutor: A idade do condutor foi um factor de tarifação do seguro automóvel durante o período a que respeitam os sinistro, aplicando-se um agravamento quando o segurado ou condutor habitual tivesse menos de 25 anos. No entanto, só se registou na base de dados o ano em que o condutor completava 25 anos, eliminando-se essa informação logo que essa idade fosse atingida. Com base nessa informação classificou-se a idade do condutor em até 22 anos, com 23 ou 24 anos e 25 ou mais anos. Dada a baixa

representatividade dos casos com menos de 25 anos, no conjunto de sinistros, suspeita-se fortemente que, no registo da informação, se tenham omitido situações em que o condutor tinha menos de 25 anos, o que a verificar-se limitará as conclusões que se possam tirar relativamente a esta variável.

Anos de Carta de Condução: Tal como a idade do condutor, o número de anos de carta foi um factor de tarifação para o qual se registou apenas o ano em que seriam concluidos os dois anos de carta, eliminando-se essa informação quando se completassem os dois anos. Este facto limita muito a medição do impacto da idade da carta sobre o custo dos sinistros, variável que se considera ser das que maior impacto tem sobre o comportamento de sinistralidade. Consideraram-se dois tipos de idade de carta: menos de dois anos e dois ou mais anos.

Concelho de Residência do Segurado: Com a utilização desta variável tenta-se captar o efeito da zona habitual de circulação do veículo sobre a sinistralidade. A informação mais adequada para utilizar no processo de estimação seria o concelho ou concelhos de circulação. Como essa variável não está disponível e é de difícil recolha, optou-se pelo concelho de residência do segurado, que se pensa ter elevada correlação com o concelho de circulação.

Dado o elevado número de concelhos existentes, foi necessário proceder à agregação destes num número limitado de zonas. Para a definição das zonas foi utilizada informação, já existente, sobre frequências de sinistralidade por concelho[2]. Definiram-se oito zonas correspondentes a quatro classes de frequência de sinistralidade quer para a região norte, quer para a região sul, mais uma zona que inclui os arquipelagos dos Açores e da Madeira.

[2]Na estimação da frequência de sinistralidade esperada por concelho, utilizou-se um estimador empírico definido como a média ponderada da frequência observada em cada concelho e da frequência observada para os concelhos com densidade populacional semelhante. A ponderação atribuida à frequência do concelho foi crescente com seu o número de unidades em risco. Recorreu-se a este estimador porque se verificou existir uma fortíssima correlação entre a frequência e a densidade populacional dos concelhos.

Refira-se que se considerou, para efeito da construção das zonas geográficas, que o norte é composto pelos distritos a norte de Coimbra e Castelo Branco, inclusive, enquanto a região sul é constituida pelos restantes distritos do continente. As oito zonas geográficas que formam o continente são apresentadas em mapa anexo.

A utilização de zonas definidas com base na frequência de ocorrência de sinistros justifica-se por se ter a expectativa de que a zona de circulação influencia muito mais o número de sinistros que a sua gravidade.

Tipo de Cliente: Esta variável identifica clientes de frotas e contratos abrangidos por protocolos. Para avaliar se este tipo de contratos tem um comportamento de sinistralidade diferente, inlcuiu-se esta informação na análise. Como existem diversas frotas e vários protocolos, classificaram-se as apólices em normais e de frotas/protocolos.

Tipo de Fraccionamento do prémio: Foram utilizados quatro tipos de fraccionamento do prémio: anual (sem fraccionamento), semestral, trimestral, e bimestral. Caso se verifique que esta variável tem impacto sobre o comportamento dos custos com sinistros, isso significará que faltam incluir no modelo algumas características influentes, cujo efeito é captado pelo tipo de fraccionamento. Essas características deverão estar relacionadas com comportamentos associados aos segurados que optam por cada um dos tipos de pagamento.

Capital de Responsabilidade Civil: Tal como a variável anterior, esta variável não influencia directamente o risco, e caso seja uma variável significativa, dever-se-á ao facto de estar a captar efeitos comportamentais dos segurados que optam por cada tipo de capital. Os contratos têm cinco tipos de capital de RC: 120, 130, 240 e 500 mil contos e capital ilimitado. Como 120 e 130 mil contos representam o capital mínimo praticado pela seguradora em diferentes períodos e como existem poucas observações com capitais

iguais ou superiores a 240 mil contos, utilizaram-se três classes de capital de RC: capitais mínimos (120 e 130 mil contos), capitais intermédios (240 e 500 mil contos) e capital ilimitado.

5.3 Resultados de uma modelização genérica

Com o intuito de estimar a distribuição do custo total dos sinistros da cobertura de CCC, condicionada pelas características do risco, reconstruiu-se o custo total dos sinistros participados, somando a franquia aos encargos suportados pela seguradora. O conjunto de observações resultante pode ser entendido como uma amostra truncada e censurada do custos com sinistros, caso se procedesse sempre à reparação do veículo. A truncagem está associada à não participação de sinistros de valor inferior à franquia e a censura está associada ao facto de, na generalidade dos casos, não se proceder à reparação de veículos quando o seu custo é superior ao valor venal do veículo.

Se se considerar que as perdas totais correspondem aos sinistros cujo custo de reparação excede o valor venal, pode-se estimar a distribuição dos custos totais associados à reparação do veículo com base na amostra truncada e censurada, recorrendo aos modelos tobit generalizados.

Na estimação da distribuição dos custos totais de reparação, consideraram-se os valores das franquias e dos veículos seguros, das apólices associadas aos sinistros, como os pontos de truncagem e de censura, respectivamente. Para a estimação dos parâmetros da distribuição, condicionados pelas variáveis exógenas, procedeu-se à maximização de funções de log-verosimilhança com a seguinte forma

$$\begin{aligned} l &= \sum_{i=1}^{n_0} \left\{ \ln\left(\int_{c_2}^{+\infty} f\left(y;\boldsymbol{\theta}_i\right) dy \right) - \ln\left(\int_{c_1}^{+\infty} f\left(y;\boldsymbol{\theta}_i\right) dy \right) \right\} + \\ &+ \sum_{i=n_0+1}^{n_0+n_1} \left\{ \ln\left(f\left(y_i;\boldsymbol{\theta}_i\right) \right) - \ln\left(\int_{c_1}^{+\infty} f\left(y;\boldsymbol{\theta}_i\right) dy \right) \right\}, \end{aligned}$$

onde a notação tem o significado que lhe foi atribuído no Capítulo 3, nomeadamente, n_0 é o número de sinistros cuja indemnização está censurada e n_1 é o número de sinistros em que o valor da indemnização é obervado sem qualquer restrição.

Na modelização ensaiaram-se 3 distribuições diferentes: a lognormal, a gama e a inversa Gaussiana (com parametrização IG2). Na modelização com distribuição gama, experimentou-se a utilização da aproximação de Peizer e Pratt, mas, pelo facto de ser uma função com dois ramos, esta gerou problemas no processo de estimação, pelo que se optou pela aproximação de Wilson-Hilferty, que não originou nenhum problema no procedimento de estimação. Utilizou-se a função de ligação logarítmica no caso das distribuições gama e inversa Gaussiana, e no caso da distribuição lognormal utilizou-se a função de ligação identidade.

Para serem utilizadas no processo de estimação, todas as variáveis exógenas qualitativas foram transformadas em variáveis artificiais. Recorde-se que a variável valor seguro, foi utilizada como variável quantitativa.

Como se referiu, nas tarifas de seguro automóvel de quase todas as seguradoras a operar no mercado português, o prémio de CCC é calculado aplicando uma taxa ao valor seguro e multiplicando o valor assim obtido por factores de desconto ou agravamento, em função de determinadas características do risco. Na definição de tarifas com esse formato, as taxas são determinadas com base no valor esperado das indemnizações agregadas[3], os descontos e agravamentos são estimados no pressuposto de uma tarifa multiplicativa e é habitual assumir-se que o valor seguro não afecta o comportamento da frequência de sinistralidade.

Nestas condições, o valor esperado do custo dos sinistros de determinado risco pode

[3]Como se viu no capítulo 4, o valor esperado das indemnizações agregadas de determinado risco corresponde à multiplicação entre os valores esperados do número e do custo dos sinistros desse risco.

ser sintetizado como

$$\mu_i = C_i \times \gamma \times (1+\omega_1)^{x_{i1}} \times ... \times (1+\omega_p)^{x_{ip}}\,, \tag{30}$$

onde, para o risco i, C é o valor seguro, γ é um factor geral, ω_j $(j = 1, ..., p)$ são factores de agravamento $(\omega_j > 0)$ ou desagravamento $(\omega_j < 0)$ do valor base $(C_i \times \gamma)$, e x_{ij} são variáveis binárias que assumem o valor 1 se o risco tem determinada característica e 0 caso contrário. A expressão (30) pode ser reescrita da seguinte forma:

$$\mu_i = \exp\left\{\varphi + \ln C_i + \sum_{j=1}^{p} x_{ij}\psi_j\right\},$$

com $\varphi = \ln(\gamma)$ e $\psi_j = \ln(1+\omega_j)$.

Estas expressões têm por hipótese básica que o valor esperado do custo é proporcional ao valor seguro. Assim, a forma tradicional de cálculo do custo médio pode ser entendida como sendo resultante de

$$\mu_i = \exp\left\{\alpha + \beta_0 \ln C_i + \sum_{j=1}^{p} x_{ij}\beta_j\right\} = C_i^{\beta_0} \exp\left\{\alpha + \sum_{j=1}^{p} x_{ij}\beta_j\right\}, \tag{31}$$

impondo-se a restrição $\beta_0 = 1$. Neste trabalho também se vai testar essa restrição, modelizando o valor esperado do custo associado aos sinistros através de (31), isto é, sem qualquer restrição sobre o parâmetro β_0.

Na amostra, observa-se que o comportamento dos custos dos sinistros próximos do valor seguro da apólice é irregular, facto que se fica a dever a diversas circunstâncias focadas anteriormente, como sejam, o valor dos salvados, os custos de peritagem e as situações em que se acorda a reparação para além o valor seguro. A conjungação destes factores leva a que sinistros de perda total tenham um custo final com comportamento aleatório,

podendo nalguns casos ser superior ao valor seguro. Como os sinistros que provocaram perdas totais não estão identificados e o código da estrada classifica como salvados os veículos cuja reparação ultrapassa 70% do seu valor venal, censuraram-se ficticiamente os custos dos sinistros em 70% do valor venal dos veículos. Este procedimento implicou a censura de 1187 observações, ou seja, de aproximadamente 6% das observações.

No processo de modelização dos custos com sinistros começou-se por incluir todas as classes das variáveis atrás referidas. Como nem todas estas classes se mostraram significativas, foram-se eliminando sucessivamente aquelas para as quais não se rejeitava a hipótese de nulidade para o parâmetro que lhe estava associado. Quando se obteve uma modelização apenas com classes de variáveis significativas, testou-se a igualdade dos parâmetros associados a diversas classes de cada variável. Quando não se rejeitou a hipótese de dois parâmetros serem iguais, introduziu-se a restrição de terem o mesmo valor. Os testes de nulidade dos parâmetros foram efectuados recorrendo às propriedades assimptóticas dos estimadores de máxima verosimilhança, enquanto nos testes de igualdade entre parâmetros se utilizou o teste da razão de verosimilhanças. Todos os testes foram realizados com uma dimensão de 5%.

Os passos descritos no parágrafo anterior não foram executados de uma forma "cega", tendo-se sempre analisado se as hipóteses que não eram rejeitadas do ponto de vista estatístico, faziam sentido do ponto de vista actuarial.

Os resultados das modelizações dos custos associados aos sinistros, após se eliminarem as variáveis não significativas e se agruparem as classes com impactos semelhantes, são apresentados na Figura 9. Na estimação foram utilizados 21071 sinistros.

Na Figura 9, o valor do termo independente corresponde à estimativa do parâmetro α, enquanto os restantes valores correspondem às estimativas dos parâmetros β_j $(j = 0, ..., p)$, da expressão (31). No caso do modelo lognormal, para facilitar a comparação das estimati-

Variável	Classe	Estimativas		
		Lognormal	Gama	Inv. Gauss.
Termo Independente		7,7300	5,7580	7,5044
Valor Seguro (em logaritmo)		0,3654	0,4985	0,3823
Sexo	Empresa	-	-	-
	Masculino	0,0430	0,0461	0,0436
	Feminino	-	-	-
Tipo de Cliente	Normal	-	-	-
	Frota	-0,3113	-0,2807	-0,3121
Fraccionamento do Prémio	Anual	-	-	-
	Semestral e Trimestral	0,0628	0,0493	0,0575
	Bimestral	0,1353	0,1258	0,1266
Capital de RC	Mínimo e Intermédio	-	-	-
	Ilimitado	-0,1422	-0,1440	-0,1262
Cilindrada	até 1500 cc	-	-	-
	de 1501 a 2500 cc	0,1118	0,1083	0,1071
	mais de 2500 cc	0,2589	0,2563	0,2500
Idade do Veículo	0 anos	-	-	-
	1 ano	-0,0917	-0,1242	-0,0850
	2 a 4 anos	-0,0744	-0,1253	-0,0778
	mais de 4 anos	-	-	-
Idade do Condutor	até 24 anos	0,1600	0,1308	0,1544
	mais de 24 anos	-	-	-
Anos de Carta	menos de 2 anos	0,2110	0,2278	0,2011
	2 ou mais anos	-	-	-
Concelho de Residência	Zona Sul 4	-	-	-
	Zona Sul 3	0,1155	0,1048	0,1142
	Zona Sul 2	0,1751	0,2026	0,1538
	Zona Sul 1	-	-	-
	Zona Norte 4	-0,0689	-0,1116	-0,0637
	Zona Norte 3 e 2	0,0586	-	0,0635
	Zona Norte 1	0,1458	0,1688	0,1379
Parâmetro de dispersão		0,8847	1,0654	0,9432

Figura 9: Estimativas dos parâmetros de modelos truncados e censurados (valores em escudos)

vas dos vários modelos, o termo independente não corresponde a $\hat{\alpha}$, mas sim a $\hat{\alpha}+\frac{\hat{\sigma}^2}{2}$. Para cada variável, existe uma classe que está integrada no termo independente e que, como tal, não possui estimativa específica. As estimativas dos parâmetros de dispersão apresentadas na Figura 9 correspondem aos parâmetros σ, m e ϕ, das expressões (16), (17) e (21), das distribuições lognormal, gama e inversa Gaussiana (na forma IG2), respectivamente.

Os resultados completos das modelizações, fornecidos pelo TSP, podem ser consultados em anexo, assim como os respectivos programas.

A análise das estimativas permite constatar que os impactos das diversas variáveis sobre o valor esperado do custo são, salvo algumas excepções, muito semelhantes nas modelizações com distribuições lognormal e inversa Gaussiana. No caso da modelização com distribuição gama, alguns parâmetros apresentam estimativas bastante diferentes daquelas que foram obtidas com as outras duas modelizações. Os casos mais significativas são as estimativas do termo independente e do parâmetro associado ao valor seguro. Este resultado é de certa forma surpreendente porque não existem razões para se supor que os resultados com o modelo gama devam ser muito diferentes daqueles que se obtêm com os modelos lognormal e inversa Gaussiana.

Este facto leva a suspeitar que existam alguns problemas de estimação com a utilização da distribuição gama. Muito provavelmente, esses problemas derivam da utilização, na função de log-verosimilhança, da aproximação de Wilson-Hilferty para a função de distribuição gama. Como se referiu no Capítulo 3, esta é uma possível fonte de enviesamento das estimativas. Nestas condições, não é aconselhável dar muita credibilidade aos resultados obtidos com o modelo gama.

Nos resultados da Figura 9 a variável idade do veículo aparece significativa apenas para idades de 1 a 4 anos, resultado que, à partida, é pouco compreensível. No entanto, existindo uma forte correlação (negativa) entre o valor seguro e a idade do veículo, este

resultado pode traduzir uma correcção ao impacto da variável valor seguro. Para avaliar o impacto da idade do veículo sobre a distribuição do custo com sinistros, efectuou-se uma modelização, com distribuição lognormal, em que se impôs a restrição de o parâmetro associado ao valor seguro ser igual a 1 ($\beta_0 = 1$ na expressão (31)). Nessas condições, os resultados mostraram que o valor esperado do custo é crescente com a idade do veículo. Os resultados desta estimação são apresentados em anexo.

Em todas as modelizações apresentadas neste trabalho, o valor seguro foi utilizado como variável explicativa. Este valor reflete o valor venal do veículo. No entanto, o valor do veículo em novo poderá ser também uma variável com poder explicativo sobre o comportamento dos custos. Quando ocorre uma perda parcial, o veículo é reparado, sendo para isso necessário adquirir peças novas. Enquanto que o valor venal de um veículo depende da sua idade, o valor de uma peça nova não se reduz por se destinar a um veículo mais velho. Assim, é natural que o valor das reparações esteja bastante correlacionado com o valor em novo do veículo. Infelizmente não foi possível utilizar esta variável, porque a base de dados não dispõe desta informação.

Um resultado da modelização dos custos associados aos sinistros que se evidencia é a não proporcionalidade entre o custo médio e o valor seguro. Enquanto que tradicionalmente se assume que o coeficiente β_0 de (31) é igual a 1, os resultados apontam para um valor ligeiramente inferior a 0,4. Em termos práticos, isto significa que, mantendo tudo o resto constante, se o valor do veículo duplicar, o custo médio está longe de duplicar.

Os resultados da Figura 9 permitem também concluir que o custo médio dos sinistros de apólices de frotas é bastante menor que os dos restantes contratos, que os segurados que optam por prémios fraccionados têm maiores custos médios, que a gravidade dos sinistros é crescente com a cilindrada do veículo e que os condutores mais novos e com menor experiência, estão associados a sinistros com maiores danos nos veículos.

Refira-se que o facto de as estimativas dos parâmetros associados às variáveis capital de responsabilidade civil e fraccionamento do prémio serem significativos indica que estas variáveis estão a captar o comportamento médio dos segurados que optam pelas suas diferentes classes. Dado que estas duas variáveis não condicionam directamente o comportamento da cobertura em análise, este fenómeno resulta da omissão de algumas variáveis.

De todas as variáveis utilizadas, a categoria, o peso bruto do veículo e o tipo de franquia foram as únicas que não se mostraram significativas.

Note-se que as estimativas apresentadas se referem apenas ao modelo dos custos, não se podendo, por isso, tirar conclusões sobre o impacto que cada variável deverá ter sobre o prémio a cobrar por cada risco. Por exemplo, verificou-se que os riscos de frotas têm um efeito de redução de cerca de 27%, sobre o valor esperado dos custos, no entanto, os riscos de frotas poderão implicar um agravamento no número de sinistros declarados suficientemente elevado para que, a combinação dos dois efeitos contrários, tenha como consequência um agravamento do prémio para este tipo de riscos.

Para avaliar a qualidade das modelizações obtidas com as três distribuições ensaiadas utilizaram-se dois critérios. O primeiro procura avaliar a qualidade global do ajustamento e baseia-se no valor da máxima log-verosimilhança, enquanto o segundo, compara o número observado e estimado de sinistros com custo superior a 70% do valor seguro (o ponto de censura utilizado na estimação), para a amostra utilizada na modelização.

Apesar de não se poder utilizar como critério o valor alcançado pela função de log-verosimilhança no seu ponto máximo, para definir qual a distribuição mais adequada, este valor pode ser indicativo da qualidade do ajustamento conseguido com cada uma das distribuições. Guiahi (2000) utiliza como critério para comparar a qualidade dos ajustamentos obtidos com várias distribuições o AIC (Akaike´s Information Criterion),

Modelo	Máxima log-verosimilhança	AIC
Lognormal	-274332	548708
Gama	-275228	550498
Inversa Gaussiana	-274293	548630

Figura 10: Qualidade do ajustamento de modelos truncados e censurados com diversas distribuições

que é definido por

$$AIC = -2\{\text{máxima log-verosimilhança} \; - \; \text{número parâmetros estimados}\}.$$

Segundo este critério, quanto menor for o valor da estatística AIC melhor o ajustamento. Na Figura 10 são apresentados os valores da máxima verosimilhança e da estatística AIC.

Qualquer dos critérios aponta para qualidades de ajustamento semelhantes para os modelos lognormal e inversa Gaussiana, e um ajustamento de pior qualidade para o modelo gama. Apesar do modelo com distribuição inversa Gaussiana apresentar os melhores valores, a diferença relativamente ao modelo lognormal não é significativa, podendo-se considerar que têm qualidade equivalente.

Para comparar os ajustamentos conseguidos, também se calculou a estimativa do número de sinistros, participados, de valor superior a 70% do valor seguro, para as três modelizações (ver Figura 11). Das três distribuições ensaiadas, a gama foi aquela cuja estimativa mais se aproximou do número de sinistros observados. No entanto, todas as distribuições fornecem estimativas significativamente diferentes do valor observado.

O número esperado de sinistros foi calculado como a soma de $\Pr(Y > 0,7L|Y > D)$ (Y, D e L com o significado atribuído nos capítulos anteriores) para todos os sinistros, dadas as suas características exógenas.

Em ambos os critérios, o modelo lognormal fornece resultados aceitáveis. Como para

	Número de Sinistros	Desvios
Observado	1187	
Lognormal	1313	-126
Inversa Gaussiana	1382	-195
Gama	1082	105

Figura 11: Número estimado de sinistros com custo superior a 70% do capital seguro

o modelo lognormal, se garante sempre a convergência para o máximo global da função de log-verosimilhança, este modelo será utilizado como referência. Assim, daqui em diante serão apresentados resultados apenas para o modelo lognormal.

5.4 Resultados de uma modelização alternativa

Na secção anterior apresentou-se uma modelização dos custos com sinistros da cobertura de CCC, tendo-se ensaiado várias distribuições. As hipóteses principais dessa modelização foram: os custos associados a cada sinistro pertencem a uma única família de distribuições; os sinistros que gerem danos inferiores à franquia nunca são declarados; as indemnizações são calculadas com base no capital seguro, sempre que o valor da reparação seja superior a esse capital.

Essa modelização foi classificada de "abordagem genérica" porque recorre a uma ferramenta que se adapta bem a custos com sinistros na presença de franquias e limites de indemnização: os modelos tobit generalizados. Ela pode facilmente transpôr-se para outros seguros, com características semelhantes no que respeita a franquias e limites de indemnização.

No entanto, se forem tidos em consideração alguns aspectos específicos dos sinistros que se pretendem modelizar, poderá não ser a abordagem mais adequada.

A modelização utilizada na secção anterior pressupõe, como se disse, que os custos com sinistros seguem a mesma família de distribuições. No caso dos sinistros da cobertura de CCC, existem alguns motivos para se suspeitar que essa hipótese pode não se verificar.

De facto, os sinistros de perda total, quer pelo tipo de acidente que lhe estão subjacentes, quer por poderem envolver o valor do salvado, poderão ter um comportamento diferente dos sinistros de perda parcial. Os resultados da Figura 11 reforçam esta suposição, pois nenhuma das distribuições forneceu uma estimativa suficientemente próxima do número de sinistros com danos superiores a 70% do valor do veículo.

Com o objectivo de ajustar a modelização à hipótese de existirem dois tipos de sinistros com naturezas diferentes, procedeu-se à estimação através de uma abordagem alternativa.

5.4.1 Pressupostos da abordagem alternativa

Considera-se que os sinistros podem ser classificados em duas categorias diferentes: sinistros de perda parcial e sinistros de perda total. Sejam Y^{PP} e Y^{PT} as variáveis aleatórias que correspondem aos custos subjacentes a sinistros de perda parcial e de perda total, respectivamente.

Dado que ocorre um sinistro, ele ou é de perda parcial, ou é de perda total. Assim, o tipo de sinistro gerado, W, é uma variável aleatória binária com distribuição de Bernoulli, isto é

$$W = \begin{cases} 0 & \text{se for sinistro de perda parcial} \\ 1 & \text{se for sinistro de perda total} \end{cases}, \tag{32}$$

sendo p a probabilidade de um sinistro ser de perda total.

Se Y for a variável aleatória que representa o custo de um sinistro, independentemente do seu tipo, então

$$Y = \begin{cases} Y^{PP} & \text{se } W = 0 \\ Y^{PT} & \text{se } W = 1 \end{cases}.$$

Variável	Classe	Estimativas
		Lognormal
Termo Independente		9,8611
Valor Seguro (em logaritmo)		0,2162
Sexo	Empresa	-
	Masculino	0,0490
	Feminino	-
Tipo de Cliente	Normal	-
	Frota	-0,2903
Fraccionamento do Prémio	Anual	-
	Semestral, Trimestral ou Bimestral	0,0858
Capital de RC	Mínimo e Intermédio	-
	Ilimitado	-0,1717
Cilindrada	até 1500 cc	-
	de 1501 a 2500 cc	0,1328
	mais de 2500 cc	0,3096
Idade do Veículo	0 anos	-
	1 ano	-0,0562
	mais de 1 ano	-
Idade do Condutor	até 24 anos	0,1848
	mais de 24 anos	-
Anos de Carta	menos de 2 anos	0,2411
	2 ou mais anos	-
Concelho de Residência	Zona Sul 4	-
	Zona Sul 3	0,1313
	Zona Sul 2	0,1702
	Zona Sul 1	-0,0925
	Zona Norte 3	0,1552
	Zona Norte 2	0,1022
	Zona Norte 1	0,1684
Parâmetro de dispersão		0,8922

Figura 12: Estimativas do modelo de custos de perdas parciais (valores em escudos)

As variáveis exógenas utilizadas na modelização das perdas parciais foram todas aquelas para as quais existia informação disponível. Tal como na "abordagem genérica", considerou-se a função de ligação identidade e recorreu-se à distribuição lognormal. As estimativas dos parâmetros, obtidas após se ter seguido a metodologia atrás indicada, são apresentadas na Figura 12. Refira-se que se utilizaram 19884 observações.

Dado que nesta estimação apenas se pretende modelizar o comportamento dos sinistros de perda parcial, o poder explicativo de uma eventual variável traduzindo o valor do veículo em novo deve ser elevado, pelas razões que já se referiram. Uma vez mais se lamenta não se dispor de tal variável.

A modelização das perdas parciais possui características muito semelhantes à da "abordagem genérica" com distribuição lognormal. Em ambas as modelizações: se utilizou a distribuição lognormal, com função de ligação identidade; a amostra utilizada possui truncagem à esquerda no valor da franquia; utilizaram-se as mesmas variáveis exógenas. Verificou-se, também, nas duas modelizações, que as observações com custos iguais ou superiores a 70% do valor seguro foram sujeitas a restrições, embora de naturezas diferentes: a censura, no primeiro caso, e a truncagem no caso em análise. Caso ambos os modelos estivessem bem especificados e tivessem sido estimados com todas as suas variáveis explicativas, como os métodos de estimação foram ajustados às restrições amostrais, as estimativas dos parâmetros deveriam ser próximas. No Capítulo 3 exemplificou-se que, forçando diversos tipos de restrições nas observações, as estimativas dos parâmetros pouco variam.

A comparação das estimativas dos parâmetros das Figuras 9 e 12 permite verificar que as diferenças nos coeficientes associados às variáveis de natureza qualitativa não são, na maioria dos casos, muito significativas. No entanto, são de assinalar algumas alterações nas estimativas dos parâmetros: agregaram-se as categorias de fraccionamento semestral e trimestral com a categoria bimestral, assumindo o parâmetro um valor intermédio aos dois parâmetros anteriores; os veículos de maior cilindrada têm, nas perdas parciais, um maior agravamento; o factor idade do veículo aparece agora mais alisado, embora continue a traduzir os problemas que já se referiram; nas zonas do concelho de residência observam-se diversas alterações, como sejam, a Zona Sul 1 passar a ser significativamente diferente da Zona Sul 4, a Zona Norte 4 deixar de ser significativa e rejeitar-se a hipótese de as Zonas Norte 3 e Norte 2 terem parâmetros iguais.

Se se comparar as estimativas do termo independente e do valor seguro obtidas com as duas modelizações, constata-se que, no modelo específico das perdas parciais, o impacto do

valor seguro sobre o valor esperado dos custos é bastante inferior àquele que se estimou na modelização conjunta de todos os sinistros. No modelo das perdas parciais, a redução do impacto do valor seguro é acompanhado pelo aumento do coeficiente associado ao termo independente.

A omissão de variáveis explicativas, porventura importantes, e a possibilidade dos custos com sinistros não seguirem distribuições lognormal contribui para a variação das estimativas, quando se passa de um modelo truncado e censurado para um modelo duplamente truncado.

Os resultados fornecem uma forte indicação de que o valor seguro é menos significativo na explicação do comportamento das perdas parciais, do que das perdas totais. Este resultado vem confirmar as expectativas iniciais de que os dois tipos de sinistros têm comportamentos diferenciados.

5.4.3 A modelização das perdas totais

Identificaram-se como sinistros de perda total, aqueles cujo custo excedeu 70% do valor seguro. Esta forma de identificação dos sinistros implica que a distribuição dos custos deste tipo de sinistros só seja válida para $Y^{PT} > 0,7L$, sendo L o valor, conhecido, do veículo seguro e Y^{PT} o custo de um sinistro de perda total.

Para modelizar os custos dos sinistros de perda total optou-se pela distribuição normal. A escolha desta distribuição resulta de não existir qualquer expectativa quanto às características da distribuição deste tipo de sinistros. Assim, outras distribuições também poderiam ser ensaiadas.

Tal como nas restantes modelizações apresentadas, assume-se que a distribuição dos custos de sinistros de perda total pode depender das características do risco.

Usualmente, quando se utiliza a distribuição normal, modeliza-se uma variável Y_i com

o valor esperado variável e variância constante, isto é, $Y \sim N\left(\mu_i, \sigma^2\right)$. Como os custos com sinistros de perda total são extremamente correlacionados com o valor seguro e a informação sugere que a amplitude dos custos é crescente com o valor seguro, concluiu-se que os custos com sinistros não são homocedásticos. Perante as evidências exibidas pelos dados, assumiu-se como padrão de heterocedasticidade, que o coeficiente de variação, $\delta = \frac{\sigma_i}{\mu_i}$, é constante, o que implica naturalmente que $\sigma_i^2 = (\delta\mu_i)^2$ é variável.

Por definição, $Y^{PT} > 0,7L$, logo, a distribuição de Y^{PT} é truncada à esquerda. Como não se considera razoável que o custo de um sinistro exceda 130% do valor seguro, o que levou a que se retirassem, da amostra utilizada, todos os sinistros nessas condições, a distribuição encontra-se também truncada à direita.

Assim, procedeu-se a estimação de Y_i^{PT}, assumindo que tem distribuição normal, reparametrizada e duplamente truncada, a que corresponde a função densidade

$$f\left(y;\mu_i,\delta\right) = \frac{\frac{1}{\delta\mu_i\sqrt{2\pi}}\exp\left\{-\frac{(y-\mu_i)^2}{2\delta^2\mu_i^2}\right\}}{\int_{0,7L_i}^{1,3L_i}\frac{1}{\delta\mu_i\sqrt{2\pi}}\exp\left\{-\frac{(y-\mu_i)^2}{2\delta^2\mu_i^2}\right\}},$$

onde μ_i é função linear das variáveis exógenas.

No processo de estimação experimentaram-se todas as variáveis explicativas disponíveis. Foram utilizadas 1187 observações de sinistros de perda total. Após se eliminarem as variáveis sem poder explicativo obtiveram-se as seguintes estimativas de máxima log-verosimilhança (valores em escudos):

$$\hat{\mu}_i = 29991 + 0,647341\,L_i \tag{34}$$

$$\hat{\delta} = 0,299998.$$

Resultados mais detalhados da estimação dos parâmetros são apresentados em anexo.

Note-se que a distribuição truncada foi ajustada para valores superiores a $0,7L_i$ e que o máximo da distribuição normal "completa" estimada é em, aproximadamente, $0,65L_i$. Assim, ao contrário do que poderia sugerir a utilização de uma distribuição normal, o comportamento do custo dos sinistros de perda total é assimétrico, uma vez que é modelizado recorrendo apenas a parte da aba direita da distribuição normal.

O resultado (34) evidencia que o valor seguro (L_i) foi a única variável que se mostrou significativa na explicação do parâmetro μ_i da distribuição normal truncada. Este é um resultado interessante, uma vez que indica que, quando ocorre um sinistros de perda total, o seu valor é influenciado quase exclusivamente pelo valor seguro. Este resultado confirma que os comportamentos dos sinistros de perda parcial e de perda total são diferentes, justificando-se que sejam modelizados separadamente.

5.4.4 A modelização da probabilidade de perda total

Quando ocorre um sinistro, ele pode ser classificado apenas de duas formas: perda total ou perda parcial. A distribuição de Bernoulli surge como indicada para modelizar o tipo de sinistro, uma vez que se trata de uma variável binária. A função probabilidade de W, a variável que identifica o tipo de sinistro e que assume os valores indicados em (32), é

$$h(w_i; p_i) = p_i^{w_i}(1-p_i)^{(1-w_i)} \qquad w = 0, 1,$$

onde o indice i pressupõe que a probabilidade de cada um dos acontecimentos varia com as características do risco. Tem-se naturalmente $p_i = \Pr(W_i = 1) = \Pr(Y_i > 0,7L_i)$.

Devido ao efeito de truncagem provocado pelas franquias, a modelização de W_i terá de ser baseada numa amostra de sinistros participados e não de sinistros ocorridos. A

probabilidade de um sinistro participado pelo risco i, ser de perda total, vem

$$p_i^o = \Pr(W_i = 1 | Y_i > D_i) = \Pr(Y_i > 0,7L_i | Y_i > D_i), \tag{35}$$

ou seja

$$\begin{aligned} p_i^o &= \frac{\Pr(Y_i > 0,7L_i)}{1 - \Pr(Y_i \leqslant D_i)} = \frac{\Pr(Y_i > 0,7L_i)}{1 - \Pr(Y \leqslant D_i | Y_i < 0,7L_i) \Pr(Y_i < 0,7L_i)} \\ &= \frac{p_i}{1 - \Pr\left(Y_i^{PP} \leqslant D_i\right)(1 - p_i)}. \end{aligned} \tag{36}$$

atendendo a que $\Pr(Y_i > 0,7L_i) = p_i$. Note-se ainda que $\Pr\left(Y_i^{PP} \leqslant D_i\right)$ pode ser estimado com base nos resultados da modelização das perdas parciais. A expressão (36) é válida para $D_i < 0,7L_i$; nos restantes casos $p_i^o = 1$.

Para modelizar o comportamento da variável W_i, ajustou-se às observações uma distribuição de Bernoulli, com parâmetro p_i^o na forma de (36), onde $\Pr\left(Y_i^{PP} \leqslant D_i\right)$ foi substituido pela sua estimativa. Considerou-se que p_i é função de uma combinação linear das variáveis exógenas, assumindo essa função a seguinte forma:

$$p_i = \frac{1}{1 + \exp\left(-\sum_{j=1}^{p} x_{ij}\beta_j\right)}. \tag{37}$$

Para estimar os parâmetros β_j, procedeu-se à maximização da função de log verosimilhança, que assume a forma

$$l = \sum_i \left[w_i \ln\left(\frac{p_i^o}{1 - p_i^o}\right) + \ln(1 - p_i^o) \right]. \tag{38}$$

Note-se que, se na expressão (38) estivesse p_i em substituição de p_i^o, este modelo seria equivalente ao modelo logit.

Na modelização, utilizaram-se as mesmas variáveis exógenas que foram utilizadas nos

Variável	Classe	Estimativas
		Bernoulli
Termo Independente		10,5849
Valor Seguro (em logaritmo)		-0,8899
Tipo de Cliente	Normal	-
	Frota	-0,8733
Fraccionamento do Prémio	Anual, Semestral ou Trimestral	-
	Bimestral	0,4562
Capital de RC	Mínimo e Intermédio	-
	Ilimitado	-0,2882
Cilindrada	até 1500 cc	-
	de 1501 a 2500 cc	0,2135
	mais de 2500 cc	0,6428
Idade do Veículo	0 anos	-
	1 ano	-0,4610
	2 ou 3 anos	-0,7541
	4 ou 5 anos	-0,8461
	6 anos	-0,5265
	7 anos	-0,3163
	8 ou mais anos	-
Idade do Condutor	até 24 anos	0,3723
	mais de 24 anos	-
Concelho de Residência	Zona Sul 4	-
	Zona Sul 3	-
	Zona Sul 2	0,2891
	Zona Sul 1	-
	Zona Norte 4 ou 3	-0,7010
	Zona Norte 2	-0,3168
	Zona Norte 1	-

Figura 13: Estimativas do modelo de probabilidade de perda total

restantes modelos e foram utilizadas 21071 observações de sinistros.

A Figura 13 contém as estimativas dos parâmetros β_j.

Uma vez que p_i é resultado de uma transformação logística da combinação linear das variáveis, o impacto das variáveis sobre p_i não pode ser interpretado directamente dos valores da Figura 13.

A análise dos resultados da Figura 13 permite tirar algumas conclusões. Como seria de esperar, quanto menor o valor do veículo, maior é a probabilidade de um sinistro ser de perda total. Mais uma vez se verifica que os sinistros gerados por segurados que optam pelo fraccionamento bimestral são os mais perigosos. Os resultados mostram que a probabilidade de perda total é crescente com a cilindrada do veículo seguro. Verifica-se

novamente que o impacto da idade do veículo não é linear. Tal como no caso das perdas parciais, os condutores mais novos provocam sinistros com maiores danos, no entanto, não se verifica que a probabilidade de perda total esteja correlacionada com a experiência do condutor.

5.5 Comparação dos resultados das duas modelizações

5.5.1 Resultados para alguns exemplos

Nas secções 5.3 e 5.4 apresentaram-se duas formas alternativas de estimar a distribuição dos custos gerados por sinistros da cobertura de CCC. Para exemplificar os resultados fornecidos pelas duas modelizações, calcularam-se valores referentes a três aspectos da distribuição dos custos: o valor esperado das indemnizações, a probabilidade de um sinistro de perda total e a probabilidade de sinistro com prejuízo inferior à franquia. No cálculo desses valores foram utilizados quatro casos com características diferentes:

- Caso 1 - veículo no valor de 3000 contos e cilindrada inferior a 1500 cm^3;
- Caso 2 - veículo no valor de 5000 contos e cilindrada inferior a 1500 cm^3;
- Caso 3 - veículo no valor de 5000 contos e cilindrada entre 1501 e 2500 cm^3;
- Caso 4 - veículo no valor de 5000 contos, cilindrada entre 1501 e 2500 cm^3, veículo com 4 anos e seguro de frota.

Para as variáveis que não foram referidas na identificação dos casos, considera-se que as suas características se encontram contempladas pelo termo independente.

Todos os valores dos exemplos foram calculados recorrendo ao *software* Mathematica 3.0.

Valor Esperado da Indemnização: Dado um risco i, com franquia dedutível D_i e valor seguro L_i, o valor esperado de uma indemnização X_i, gerada por um dano Y_i associado a um sinistro ocorrido (tenha ou não sido participado) é, no caso da modelização genérica,

$$E(X_i) = \left[\int_{D_i}^{L_i} (y - D_i) f(y; \boldsymbol{\theta}_i) \, dy + (L_i - D_i) \int_{L_i}^{+\infty} f(y; \boldsymbol{\theta}_i) \, dy \right],$$

enquanto que, na modelização bi-etápica é, recorrendo ao teorema da probabilidade total,

$$E(X_i) = E(X_i | W_i = 0) \, . \Pr(W_i = 0) + E(X_i | W_i = 1) \, . \Pr(W_i = 1),$$

com

$$\begin{aligned} \Pr(W_i = 0) &= 1 - p_i \\ \Pr(W_i = 1) &= p_i \end{aligned}$$

$$E(X_i | W_i = 0) = \frac{\int_{D_i}^{0,7L_i} (y - D_i) \, h(y; \boldsymbol{\theta}_i') \, dy}{\int_0^{0,7L_i} h(y; \boldsymbol{\theta}_i') \, dy}$$

$$E(X_i | W_i = 1) = \frac{\int_{0,7L_i}^{1,3L_i} (y - D_i) \, g(y; \boldsymbol{\theta}_i'') \, dy}{\int_{0,7L_i}^{1,3L_i} g(y; \boldsymbol{\theta}_i'') \, dy},$$

onde $f(y; \boldsymbol{\theta}_i)$, $h(y; \boldsymbol{\theta}_i')$, e $g(y; \boldsymbol{\theta}_i'')$, são as funções densidade dos custos de todos os sinistros, dos sinistros de perda parcial e dos sinistros de perda parcial, respectivamente.

Os valores da Figura 14 mostram que, para os três primeiros casos calculados, se obtêm valores esperados superiores através da modelização bi-etápica, enquanto que no caso 4 os valores esperados resultantes das duas metodologias são semelhantes.

Franquia	Modelização	Caso 1	Caso 2	Caso 3	Caso 4
0	Genérica	519679	633125	705987	483519
	Bi-etápica	567949	667144	762524	481278
100000	Genérica	421854	534482	606992	386200
	Bi-etápica	470665	569261	664036	384429
250000	Genérica	303174	407793	476215	271760
	Bi-etápica	355469	449281	538309	273114
500000	Genérica	181221	265399	321873	159372
	Bi-etápica	237987	321673	396015	166260

Figura 14: Valores esperados das indemnizações resultantes das duas modelizações (valores em escudos)

Franquia	Abordagem	Caso 1	Caso 2	Caso 3	Caso 4
100000	Genérica	0,07473	0,04922	0,03764	0,08966
	Bi-etápica	0,08966	0,07221	0,05364	0,10255
250000	Genérica	0,34243	0,26870	0,22871	0,37938
	Bi-etápica	0,36822	0,32750	0,27480	0,40398
500000	Genérica	0,64718	0,56622	0,51611	0,68310
	Bi-etápica	0,65514	0,61882	0,56004	0,70112

Figura 15: Probabilidades de custo com sinistro inferior à franquia

Probabilidade de sinistro com prejuízo inferior à franquia: A forma de cálculo desta probabilidade é diferenciada nas duas abordagens. Na modelização genérica

$$\Pr\left(X_i = 0\right) = \Pr\left(Y_i \leqslant D_i\right) = \int_0^{D_i} f\left(y; \boldsymbol{\theta}_i\right) dy,$$

enquanto que, na modelização bi-etápica

$$\begin{aligned} \Pr\left(Y_i \leqslant D_i\right) &= \Pr\left(Y_i^{PP} \leqslant D_i\right) \Pr\left(W_i = 0\right) \\ &= \frac{\int_0^{D_i} h\left(y; \boldsymbol{\theta}_i'\right) dy}{\int_0^{0,7L_i} h\left(y; \boldsymbol{\theta}_i'\right) dy} \left(1 - p_i\right). \end{aligned}$$

Para os casos exemplificados na Figura 15, a probabilidade de um sinistro gerar um prejuizo inferior à franquia não é substancialmente diferente nas duas modelizações.

Abordagem	Caso 1	Caso 2	Caso 3	Caso 4
Genérica	0,02276	0,00898	0,01255	0,00373
Bi-etápica	0,06369	0,04139	0,05074	0,00949

Figura 16: Probabilidades de perda total

Probabilidade de Perda Total: Enquanto que na "abordagem específica" a probabilidade de perda total é $\Pr(W_i = 1) = p_i$, na "abordagem genérica" resulta de

$$\Pr(Y_i > 0,7L_i) = \int_{0,7L_i}^{+\infty} f(y;\boldsymbol{\theta}_i)\, dy.$$

Como se pode verificar na Figura 16, para os casos exeplificados, as probalidades de perda total resultantes da abordagem bi-etápica são significativamente superiores às da abordagem genérica.

5.5.2 Avaliação da qualidade dos ajustamentos

Na secção anterior apresentaram-se diversos resultados relativos às modelizações "genérica" e bi-etápica. No entanto, nenhum desses resultados permite concluir sobre qual das modelizações melhor capta o comportamento dos custos com sinistros.

Para avaliar a qualidade dos ajustamentos conseguidos com as duas modelizações, calculou-se, dadas as características exógenas dos sinistros contidos na amostra que serviu de base às modelizações, o número estimado de sinistros participados, em diversos intervalos de custo.

Os limites dos intervalos de custos com sinistros foram definidos em proporção do capital seguro. A estimativa do número de sinistros participados com custo inferior uma proporção α do capital seguro, foi obtida como a soma das probabilidades do custo de

cada sinistro participado ser inferior a essa proporção, ou seja,

$$\hat{n}(\alpha) = \sum_i \Pr(Y_i < \alpha L_i | Y_i > D_i).$$

No caso da modelização genérica, tem-se

$$\Pr(Y_i < \alpha L_i | Y_i > D_i) = \frac{\Pr(Y_i < \alpha L_i) - \Pr(Y_i < D_i)}{1 - \Pr(Y_i < D_i)},$$

enquanto que, no caso da modelização bi-etápica, vem

$$\Pr(Y_i < \alpha L_i | Y_i > D_i) = \begin{cases} \frac{\Pr\left(Y_i^{PP} < \alpha L_i\right)(1-p_i) - \Pr\left(Y_i^{PP} < D_i\right)(1-p_i)}{1 - \Pr\left(Y_i^{PP} < D_i\right)(1-p_i)} & \text{se } \alpha \leqslant 0,7 \\ 1 - \frac{\Pr\left(Y_i^{Pt} > \alpha L_i\right)p_i}{1 - \Pr\left(Y_i^{PP} < D_i\right)(1-p_i)} & \text{se } \alpha > 0,7 \end{cases}.$$

Estas expressões são válidas apenas para $\alpha L_i > D_i$, porque nos restantes casos

$\Pr(Y_i < \alpha L_i | Y_i > D_i) = 0.$

Na Figura 17 apresentam-se as estimativas do número de sinistros por intervalo do custo em proporção do valor seguro, bem como o número observado de sinistros participados em cada um destes intervalos.

A análise destes resultados permite constatar que nos primeiros intervalos ambas as metodologias são algo imprecisas. A partir dos intervalos em que o custo representa mais de 40% do valor seguro, a modelização bi-etápica fornece sempre estimativas bastante mais próximas dos valores observados, que a modelização genérica. Verifica-se que a modelização genérica não modeliza bem os sinistros de perda total, pois os desvios nos intervalos referentes a esse tipo de sinistros são elevados.

Refira-se que parte dos desvios que se verificam nos primeiros intervalos se podem dever ao fenómeno que se denomina por *sede de bonus*, e que, basicamente, consiste no facto de os segurados não participarem sinistros de pequeno montante com o objectivo de

Intervalo de	Número de sinistros			Desvios quadráticos relativos	
capital	observado	modelo genérico	modelo bi-etápico	modelo genérico	modelo bi-etápico
0% - 2,5%	213	280,65	312,52	21,48	46,50
2,5% - 5%	2035	2009,38	2069,40	0,32	0,58
5% - 10%	4927	4617,20	4521,95	19,48	33,30
10% - 15%	3347	3477,96	3366,97	5,12	0,12
15% - 20%	2336	2453,86	2397,36	5,95	1,61
20% - 25%	1632	1781,49	1767,40	13,69	11,23
25% - 30%	1190	1299,25	1312,76	10,03	12,66
30% - 35%	898	971,30	1000,40	5,98	11,68
35% - 40%	730	740,50	777,67	0,15	3,11
40% - 45%	609	575,19	615,84	1,88	0,08
45% - 50%	527	453,75	495,08	10,18	1,93
50% - 55%	458	363,05	403,49	19,68	6,49
55% - 60%	403	294,01	332,67	29,47	12,27
60% - 65%	326	240,79	277,24	22,27	7,29
65% - 70%	253	199,16	233,27	11,46	1,54
70% - 75%	245	166,15	243,28	25,38	0,01
75% - 80%	233	139,72	226,89	37,35	0,16
80% - 85%	171	118,39	197,94	16,18	4,24
85% - 90%	153	100,95	161,63	17,71	0,49
90% - 95%	138	86,59	123,61	19,15	1,50
>95%	247	701,66	233,65	836,91	0,72

Figura 17: Número estimado de sinistros por intervalo de custo

não perderem bonificações no prémio, por ausência de sinistralidade. O efeito da sede de bonus, embora decrescente com o valor do sinistro, não afecta apenas o comportamento do número de participações de pequenos sinistros, uma vez que, por exemplo, para uma franquia de 20% do valor seguro, o segurado pode não declarar um sinistro que represente 22% desse valor.

Para comparar a qualidade dos dois ajustamentos calculou-se a estatística do teste de ajustamento do χ^2. Os resultados foram: 674,23 e 146,85, para as modelizações genérica e bi-etápica, respectivamente, o que vem confirmar que a segunda abordagem permite uma melhor modelização do comportamento dos custos com sinistros. Apesar de melhor, a modelização bi-etápica não capta, ainda assim, aspectos essenciais do fenómeno.

6 A interacção entre frequência e custo

Nos capítulos anteriores focaram-se principalmente aspectos relacionados com a modelização da distribuição dos custos com sinistros. Como se expôs no Capítulo 4 e se formalizou na expressão (26), o valor esperado das indemnizações geradas por um risco resulta, assumindo determinadas hipóteses, da multiplicação entre os valores esperados do custo de cada sinistro e do número de sinistros. Assim, se se tiver como objectivo a estimação de um estrutura tarifária, para além de se modelizar a distribuição dos custos, deve-se também analisar o comportamento da frequência de sinistralidade dos diversos riscos, tendo em conta as suas características.

As franquias e os limites de indemnização, as duas características que conduziram a que se propusesse uma metodologia alternativa aos modelos lineares generalizados, têm impactos diferenciados sobre a frequência de sinistralidade. Dado que todos os sinistros, cujos custos se aproximem ou excedam os limites de indemnização, são participados, esses limites não têm qualquer impacto sobre o número de sinistros declarados à seguradora. Já as franquias, por seu lado, afectam o número de sinistros declarados, uma vez que os segurados não têm qualquer benefício em declarar os sinistros cujo montante seja inferior à franquia.

6.1 A estimação da frequência

Uma vez que o comportamento do número de sinistros participados à seguradora, por um determinado risco, depende da franquia a que este está sujeito, é necessário que, na modelização da frequência de sinistralidade, se considerem os montantes das franquias associadas aos riscos.

Recorrendo à metodologia aplicada no capítulo anterior, estima-se o comportamento do custo de um sinistro, na ausência de franquias, isto é, numa situação em que todos os

sinistros ocorridos seriam participados.

Numa carteira com franquias, o que se conhece são os sinistros participados, que são em menor quantidade do que os ocorridos. Assim, para se construir uma tarifa, é necessário estimar a frequência de sinistralidade que se verificaria na ausência de franquias.

Seja $f(y;\boldsymbol{\theta}_i)$ a função densidade de probabilidade dos prejuízos gerados por um sinistro, para uma determinada apólice. Se essa apólice tiver franquia, de valor D_i, os sinistros de valor inferior a D_i não serão participados. Assim, quando ocorre um sinistro, existe uma probabilidade de ele não ser declarado, a qual corresponde a

$$\Pr[Y_i \leqslant D_i] = \int_0^{D_i} f(y;\boldsymbol{\theta}_i)\, dy, \tag{39}$$

valor que se pode estimar recorrendo à modelização apresentada no capítulo anterior.

Recorrendo a um teorema bem conhecido, sabe-se que, se N tem distribuição de Poisson, $N \sim P(\mu)$, e a distribuição da variável N^*, condicionada por $N = n$, é binomial $B(n;p)$, então $N^* \sim P(\lambda)$, com $\lambda = \mu p$. Este resultado pode ser encontrado, por exemplo, em Murteira (1990).

Utilizando o resultado anterior, se N_i for o número de sinistros ocorridos para um risco i, com distribuição de Poisson de parâmetro μ_i, então N_i^*, o número de sinistros participados, segue também uma distribuição de Poisson, mas com parâmetro

$$\lambda_i = \mu_i \left(1 - \int_0^{D_i} f(y;\boldsymbol{\theta}_i)\, dy\right), \tag{40}$$

uma vez que um sinistro é participado com probabilidade $p_i = 1 - \int_0^{D_i} f(y;\boldsymbol{\theta}_i)\, dy$.

Assim, utilizando (40), pode-se estimar o número esperado de sinistros participados, para cada risco individual, desde que se conheçam as distribuições dos custos e do número de sinistros ocorridos.

Habitualmente, quando se pretende modelizar o número de sinistros participados por um risco, utiliza-se a distribuição de Poisson, com parâmetro λ_i. É usual considerar-se que o parâmetro da distribuição é função da combinação linear das variáveis, com função de ligação logarítmica, ou seja,

$$\lambda_i = \exp\left(\sum_{j=1}^{p} x_{ij}\beta_j\right).$$

No entanto, essa modelização não tem em consideração os impactos das diferentes franquias dos contratos. Para ultrapassar essa insuficiência, propõe-se que se modelize o número de sinistros participados com a distribuição de Poisson, mas formalizando o parâmetro λ_i de forma diferente. Considerando que o parâmetro μ_i, que caracteriza a distribuição dos sinistros ocorridos, depende do valor das variáveis explicativas da seguinte forma

$$\mu_i = \exp\left(\sum_{j=1}^{p} x_{ij}\beta'_j\right) \tag{41}$$

então, conciliando (40) e (41), o número de sinistros participados pode ser modelizado com parâmetro

$$\lambda_i = \exp\left(\sum_{j=1}^{p} x_{ij}\beta'_j\right)\left(1 - \int_0^{D_i} f\left(y;\boldsymbol{\theta}_i\right) dy\right).$$

Na estimação de λ_i, substitui-se a quantidade $\left(1 - \int_0^{D_i} f\left(y;\boldsymbol{\theta}_i\right) dy\right)$ pela sua estimativa $\left(1 - \int_0^{D_i} f\left(y;\hat{\boldsymbol{\theta}}_i\right) dy\right)$, sendo necessário modelizar previamente os custos. Esta estimativa pode ser calculada para todos os riscos, tenham ou não gerado sinistros. Esta modelização permite "corrigir" o número esperado de sinistros do efeito da franquia, permitindo também estimar o parâmetro da distribuição do número de sinistros ocorridos que teriam

sido participados caso não existisse franquia. A estimação dos parâmetros β'_j pode ser efectuada com o método da máxima verosimilhança.

6.2 A combinação da frequência e do custo

Após se terem estimado os comportamentos do número de sinistros e do custo individual de cada sinistro, tem-se informação suficiente para se poderem construir estruturas tarifárias, para produtos com diversas características de franquia e limite de indemnização.

A expressão (26) traduz que, $E(Z)$, o valor esperado das indemnizações agregadas geradas por um risco, em determinado periodo, é igual ao produto do valor esperado do número de sinistros participados e do valor esperado de cada indemnização. Como no âmbito dos modelos lineares generalizados se modeliza directamente estes valores esperados, a aplicação da expressão (26) é imediata.

O mesmo não acontece quando se recorre à metodologia baseada nos modelos tobit generalizados, por nesta se modelizarem os custos subjacentes às indemnizações e o número de sinistros ocorridos.

Considere-se que

$$X(D,L) = \begin{cases} 0 & Y < D \\ Y - D & D \leqslant Y < L \\ L - D & Y \geqslant L \end{cases} \tag{42}$$

onde X representa o valor da indemnização paga pela seguradora em consequência de um sinistro que gera um custo de valor Y, sujeita a uma franquia dedutível de valor D e a um limite de indemnização L.

Note-se que $X(0,+\infty) = Y$, situação em que não existe franquia nem limite de indemnização.

Considere-se também que N e N^* são variáveis aleatórias que representam o número de sinistros ocorridos e o número de sinistros participados, respectivamente. Como só são participados os sinistros com custo superior à franquia, o valor esperado das indemnizações conhecidas pela seguradora é $E[X(D,L)|Y>D]$ e o valor esperado do número de sinistros participados é $E(N^*|Y>D)$. Adaptando a expressão (26), tem-se

$$E[Z(D,L)|Y>D] = E(N^*|Y>D) \times E[X(D,L)|Y>D] \tag{43}$$

onde $Z(D,L)$ é o valor das indemnizações agregadas de determinado risco.

Assim, no âmbito dos modelos tobit generalizados, para se proceder ao cálculo de prémios, com base no critério do valor esperado, deve-se utilizar a expressão (43).

A expressão (43) pode ser simplificada. Se se considerar que N segue uma distribuição de Poisson, então,

$$\begin{aligned} E[Z(D,L)|Y>D] &= E(N^*|Y>D) \times E[X(D,L)|Y>D] \\ &= E(N)\Pr(Y>D) \times E[X(D,L)] \frac{1}{\Pr(Y>D)} \\ &= E(N) \times E[X(D,L)] = E[Z(D,L)]. \end{aligned} \tag{44}$$

7 Conclusões e comentários finais

Os modelos lineares generalizados são uma das "ferramentas" mais utilizadas para modelizar as indemnizações agregadas geradas por um risco. Estes modelos permitem estimar quer o valor esperado das indemnizações individuais a cargo da seguradora, quer o valor esperado do número de sinistros participados. No entanto, não permitem modelizar adequadamente as indemnizações individuais de seguros em que existam limites de indemnização e franquias. Apesar da sua pouca adequação, os modelos lineares generalizados são utilizados frequentemente em tais situações.

Neste trabalho propôs-se uma metodologia alternativa, os modelos tobit generalizados, que permite estimar o comportamento dos prejuízos subjacentes às indemnizações, com base em indemnizações que resultam da aplicação de franquias e de limites de indemnização.

Enquanto nos modelos lineares generalizados se estima directamente o valor esperado das indemnizações, que resultaram da aplicação de determinadas franquias e limites de indemnização, nos modelos tobit generalizados utiliza-se um procedimento diferente: começa-se por estimar os parâmetros da distribuição dos prejuízos que originaram as indemnizações, após o que se calcula a estimativa do valor esperado das indemnizações individuais. Este valor esperado depende do tipo e valor da franquia, assim como, do limite de indemnização.

O procedimento descrito permite, não só, estimar as indemnizações esperadas associadas às actuais características dos contratos, como também as indemnizações associadas a qualquer outro valor e tipo de franquia e/ou limite de indemnização. Esta versatilidade dos modelos tobit generalizados representa uma das suas principais vantagens, pois facilita a criação de produtos com novas características.

A aplicação prática dos modelos tobit generalizados apresenta algumas limitações. Nestes modelos, ao contrário do que acontece nos modelos lineares generalizados quando se opta por uma função de ligação logarítmica, a tarifa, de um produto com franquias, nunca tem estrutura multiplicativa. Isto porque, nos modelos tobit generalizados, não se estima directamente o valor esperado das indemnizações, mas sim o valor esperado dos prejuízos que lhes dão origem. Esta característica dos modelos tobit generalizados é inconveniente quando o objectivo for a construção de uma estrutura tarifária simples, situação em que será necessário proceder a algumas aproximações.

Para ilustrar a utilização dos modelos tobit generalizados, estimou-se a distribuição dos custos gerados por cada sinistro da cobertura de Choque, Colisão ou Capotamento (CCC), do seguro automóvel. Apesar da insuficiente qualidade e quantidade da informação disponível, obtiveram-se alguns resultados interessantes.

Na modelização dos custos com sinistros utilizaram-se duas abordagens. Numa primeira abordagem considerou-se que os custos de todos os sinistros seguem uma mesma distribuição. Esta abordagem foi designada "genérica" por ser aplicável à modelização dos custos de diversos seguros. Ensaiaram-se três distribuições (lognormal, gama e inversa Gaussiana), tendo os resultados sugerido que a distribuição lognormal foi aquela que permitiu um melhor ajustamento. A modelização com distribuição gama forneceu estimativas significativamente diferentes das restantes modelizações, o que se suspeita que tenha origem na utilização de uma aproximação da função de distribuição insuficientemente precisa.

Na segunda abordagem, que se designou por bi-etápica, assumiu-se que existem dois tipos de sinistros com distribuições distintas: sinistros de perda parcial e sinistros de perda total. Nesta abordagem modelizaram-se separadamente as distribuições dos custos de cada um dos tipos de sinistro e a distribuição da probabilidade de um sinistro ser de cada tipo.

Verificou-se que, enquanto o comportamento dos custos das perdas parciais dependem de diversos factores, os custos dos sinistros de perda total apenas dependem do valor seguro.

A comparação dos resultados obtidos com as duas abordagens permitiu constatar que o ajustamento conseguido através da abordagem bi-etápica é de melhor qualidade.

Ao contrário do que é assumido na construção das tarifas da cobertura de CCC, o valor esperado da indemnização não é proporcional ao valor do veículo. Verificou-se que o valor marginal da indemnização esperada é decrescente com o valor do veículo. As estimativas dos parâmetros confirmam algumas expectativas sobre o comportamento dos riscos, como seja a maior gravidade dos sinistros dos condutores jovens e/ou com pouca experiência e dos veículos de maior cilindrada.

Refira-se que, ao se proporem estes modelos, não se pretende explicar o comportamento dos custos gerados pelos sinistros. Com esta metodologia procura-se apenas uma ferramenta que conduza a uma melhor aproximação das responsabilidades das seguradoras, evitando algumas imprecisões desnecessárias, e fornecer algum suporte estatístico para estimar o valor esperado das indemnizações quando se alteram os valores das franquias e/ou dos limites de indemnização.

Para que os modelos propostos possam ser utilizados com frequência, nas seguradoras, como ferramenta para a construção de estruturas tarifárias, é desejável ter outras formas, mais eficientes, de proceder à sua estimação. O TSP, por ser um software flexível, foi considerado adequado para estimar os modelos necessários para o desenvolvimento deste trabalho. No entanto, o seu processo de estimação é também pouco eficiente. Sempre que se considerou um modelo com bastantes variáveis explicativas (mais de 30) e uma distribuição diferente da lognormal, o processo de estimação foi bastante demorado. Note-se que foram utilizadas pouco mais do que 20000 observações, o que é relativamente pouco, tendo em conta a informação que por vezes é necessário modelizar, no âmbito da actividade

seguradora.

8 Bibliografia

Amemiya, T. (1973), Regression Analysis When the Dependent Variable is Truncated Normal, *Econometrica*, 42, pp. 999-1012.

Associação Portuguesa de Seguradores (2001), *Produção de Seguro Directo: 1990 a 2000*, Associação Portuguesa de Seguradores.

Barata, J.C. (2000), *Construção de uma tarifa de seguro de responsabilidade civil automóvel*, Tese de mestrado em Ciências Actuariais, ISEG, UTL.

Berg, P. Ter (1994), Deductibles and The Inverse Gaussian Distribution, *Astin Bulletin*, 24, N°2, pp. 319-323.

Centeno, L. (2000), *Teoria do Risco*, Cemapre, ISEG, Lisboa.

Dobson, A. (1990), *An Introduction to Generalized Linear Models*, London: Capman Hall.

Guiahi, F. (2000), *Fitting to Loss Distributions with Emphasis on Rating Variables*, http://www.casact.org/pubs/forum/01wforum/01wf133.pdf.

Hall, B. H., Cummins, C. e Schnake, R. (1991), *TSP Reference Manual - Version 4.2*, TSP International, Palo Alto, USA.

Hall, B. H., Cummins, C. e Schnake, R. (1991), *TSP User's Guide - Version 4.2*, TSP International, Palo Alto, USA.

Hogg, R. V. e Klugman, S. A. (1984), *Loss Distributions*, New York: John Wiley & Sons.

Johnson, N.L., Kotz, S. e Kemp, A.W. (1992), *Univariate Discrete Distributions*, Second Edition, New York: John Wiley & Sons.

Johnson, N.L., Kotz, S. e Balakrishnan, N. (1999), *Continuous Univariate Distributions - Volume 1*, Second Edition, New York: John Wiley & Sons.

Judge, G. G., Hill, R. C., Griffiths, W. E., Lütkepohl, H. e Lee, T. C. (1988), *Introduction to the Theory and Practice of Econometrics*, Second Edition, New York: John Wiley & Sons.

Ling, R.F. (1977), On maximum absolute error of some approximations for t, chi-squared, and F tail probabilities, *Asa Proceedings on Statistical Computing*, pp. 299-304.

Ling, R.F. (1978), A study of accuracy of some approximatinos for t, χ^2 and F tail probabilities, *Journal of the American Statistical Association*, 73, pp. 274-283.

Maddala, G.S. (1983), *Limited-dependent and Qualitative Variables in Econometrics*, Cambridge: Cambridge University Press.

McCullagh, P. e Nelder, J. (1989), *Generalized Linear Models*, 2nd Edition, London: Chapman & Hall.

Murteira, B. J. F. (1990), *Probabilidades e Estatística - Volume 1*, 2ªEdição, Lisboa: McGraw-Hill.

Nelder, J. A. e Wedderburn, R.W. (1972), Generalized linear models, *Journal of Royal Statistical Society*, A, 135, pp. 370-384.

Olsen, R.J. (1978), Note on the Uniqueness of the maximum Likelihood Estimator for Tobit Model, *Econometrica*, 46, pp. 1211-1215.

Tobin, J. (1958), Estimation of Relationships for Limited Dependent Variables, *Econometrica*, 26, pp. 24-36.

9 Anexos

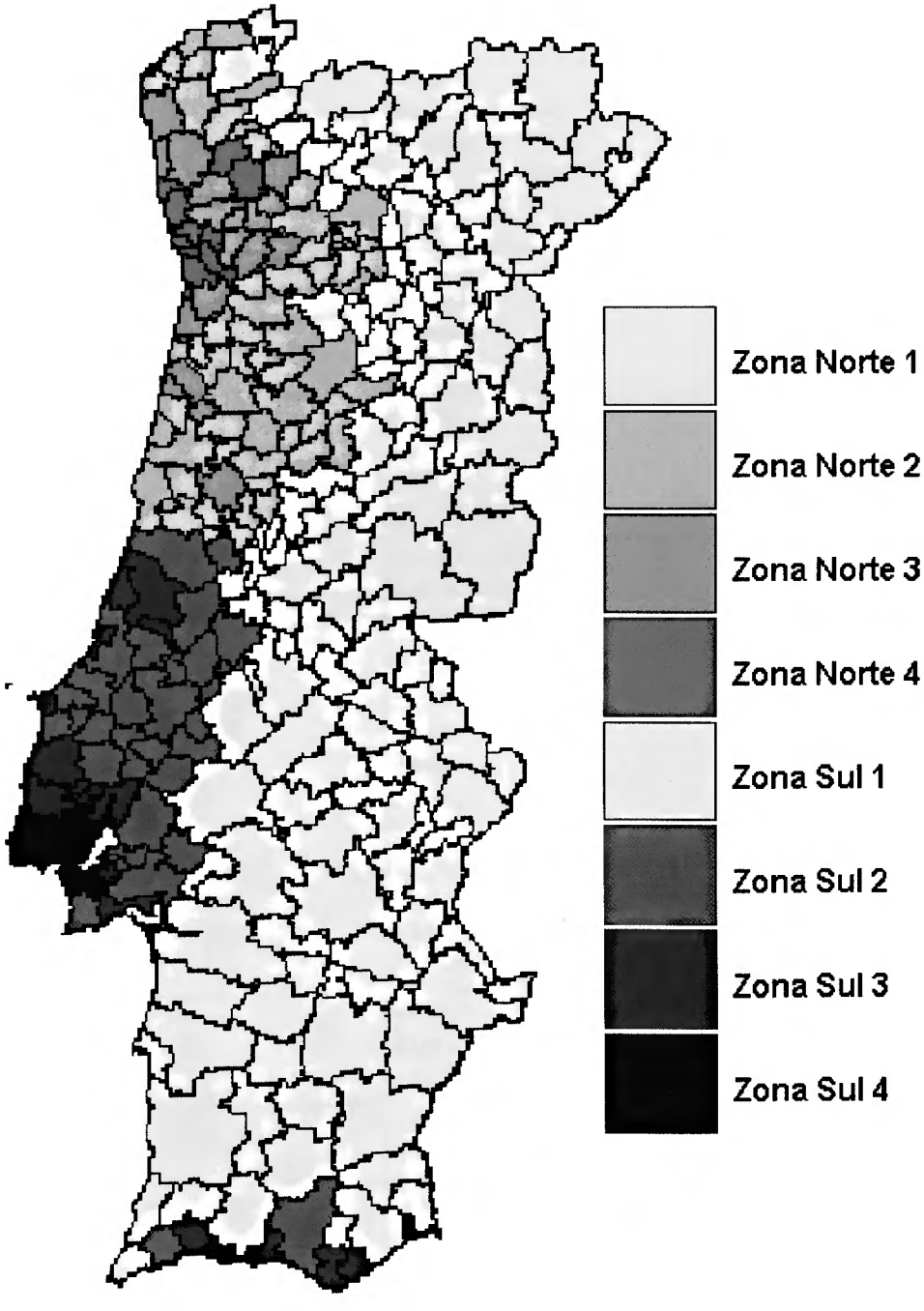
- Algumas estatísticas dos sinistros por classe das principais variáveis exógenas
- Mapa das zonas geográficas
- Identificação dos parâmetros associados às variáveis exógenas utilizados no TSP
- Resultados das modelizações referentes à abordagem "genérica": distribuições lognormal, gama e inversa Gaussiana
- Resultados da modelização dos custos com sinistros com restrição sobre o parâmetro do valor seguro
- Resultados das modelizações do custo com sinistros de perda parcial e da probabilidade de perda total (referentes à abordagem bi-etápica)
- Resultados da modelização do custo com sinistros de perda total (referente à abordagem bi-etápica)

Algumas estatísticas dos sinistros por classe das principais variáveis exógenas

Variável	Classe	Peso	Nº de sinistros	Custo médio	Franquia média	Valor Seguro médio
Sexo	Empresa	50,65%	10 673	465 616	76 354	2 673 162
	Masculino	36,71%	7 736	476 484	56 682	2 180 107
	Feminino	12,63%	2 662	433 180	47 961	1 964 029
Tipo de Cliente	Normal	80,54%	16 970	477 996	59 704	2 324 264
	Frota	19,46%	4 101	413 833	89 715	2 726 517
Fraccionamento do Prémio	Anual	69,22%	14 586	462 510	63 339	2 402 674
	Semestral	19,49%	4 106	474 672	57 689	2 197 367
	Trimestral	1,90%	401	511 453	66 369	2 537 860
	Bimestral	9,39%	1 978	459 278	97 949	2 800 164
Capital de RC	Mínimo	28,46%	5 997	464 530	49 200	1 939 945
	Intermédio	3,26%	687	511 460	58 799	2 173 181
	Ilimitado	68,28%	14 387	463 721	72 680	2 606 338
Cilindrada	até 1500 cc	46,56%	9 810	365 712	48 400	1 754 797
	de 1501 a 2500 cc	50,22%	10 582	537 194	77 862	2 869 433
	mais de 2500 cc	3,22%	679	790 139	121 289	4 485 001
Idade do Veículo	0 anos	10,69%	2 253	649 298	85 103	3 434 750
	1 ano	22,63%	4 768	550 265	84 551	3 239 665
	2 anos	19,01%	4 005	490 271	75 536	2 726 758
	3 anos	13,97%	2 944	436 468	66 347	2 263 759
	4 anos	9,76%	2 057	383 079	52 959	1 832 003
	5 anos	6,82%	1 437	369 972	39 282	1 493 573
	6 anos	5,35%	1 127	352 870	34 302	1 326 768
	7 anos	4,24%	893	321 188	33 217	1 084 533
	8 anos	2,46%	518	309 173	30 021	950 581
	9 anos	4,74%	999	275 474	34 141	810 717
	mais de 9 anos	0,33%	70	488 492	71 550	2 472 299
Idade do Condutor	até 22 anos	1,10%	232	665 246	59 161	2 111 496
	23 ou 24 anos	1,40%	294	488 818	59 810	2 175 692
	mais de 24 anos	97,50%	20 545	462 919	65 699	2 409 087
Anos de Carta	menos de 2 anos	1,40%	296	592 572	51 107	1 927 412
	2 ou mais anos	98,60%	20 775	463 698	65 751	2 409 323
Concelho de Residência	Zona Sul 4	50,63%	10 669	431 358	67 479	2 371 892
	Zona Sul 3	5,14%	1 083	538 085	60 216	2 324 175
	Zona Sul 2	5,49%	1 157	633 929	68 448	2 605 042
	Zona Sul 1	6,59%	1 389	520 963	59 754	2 400 590
	Zona Norte 4	17,66%	3 722	421 838	61 800	2 371 583
	Zona Norte 3	6,69%	1 410	498 640	70 938	2 583 999
	Zona Norte 2	4,70%	991	510 168	60 472	2 375 945
	Zona Norte 1	3,08%	650	596 927	67 356	2 504 499

(valores em escudos)

Mapa das Zonas Geográficas



Identificação dos parâmetros associados às variáveis exógenas utilizados no TSP

Variável	Classe	Parâmetro
Termo Independente		c0
Valor Seguro (em logaritmo)		ccap
Sexo	Empresa	-
	Masculino	csex1
	Feminino	csex2
Tipo de Cliente	Normal	-
	Frota	ctc1
Fraccionamento do Prémio	Anual	-
	Semestral	cfr1
	Trimestral	cfr2
	Bimestral	cfr3
Capital de RC	Mínimo	-
	Intermédio	ccrc1
	Ilimitado	ccrc2
Cilindrada	até 1500 cc	-
	de 1501 a 2500 cc	ccc1
	mais de 2500 cc	ccc2
Idade do Veículo	0 anos	-
	1 ano	civ1
	2 anos	civ2
	3 anos	civ3
	4 anos	civ4
	5 anos	civ5
	6 anos	civ6
	7 anos	civ7
	8 anos	civ8
	9 anos	civ9
	mais de 9 anos	civ10
Idade do Condutor	até 22 anos	cic1
	23 ou 24 anos	cic2
	mais de 24 anos	-
Anos de Carta	menos de 2 anos	cac1
	2 ou mais anos	-
Concelho de Residência	Zona Sul 4	-
	Zona Sul 3	cs3
	Zona Sul 2	cs2
	Zona Sul 1	cs1
	Zona Norte 4	cn4
	Zona Norte 3	cn3
	Zona Norte 2	cn2
	Zona Norte 1	cn1
Parâmetro de dispersão	distrib. lognormal	sigma
	distr. Gama	m
	distr. Inv. Gaussiana	fi

Resultados das modelizações referentes à abordagem "genérica": distribuições lognormal, gama e inversa Gaussiana

```
                                TSP Version 4.4
                            (11/18/97) DOS/Win      65MB
                        Copyright (C) 1997 TSP International
                                ALL RIGHTS RESERVED
                                 10/27/01 12:52PM
                    In case of questions or problems, see your local TSP
                    consultant or send a description of the problem and the
                    associated TSP output to:
                                   TSP International
                               P.O. Box 61015, Station A
                                  Palo Alto, CA 94306
                                         USA

          PROGRAM
LINE  ***************************************************************
1   options crt memory=65;
2   freq n;
3
3   read(file='autozona.txt ') Custo F AS1 AS2 Sex1 Sex2 Cat1 Cat2 Cat3
    Cat4 Cat5 TC1 Fr1 Fr2 Fr3 B1 B2 B3 B4 B5 B6 B7 CRC1 CRC2 PB1 PB2 PB3 CC1
    CC2 IV1 IV2 IV3 IV4 IV5 IV6 IV7 IV8 IV9 IV10 IC1 IC2 AC1 CAPRISC Z1 Z2 Z3
    Z4 Z5 Z6 Z7 Z8 S3 S2 S1 N4 N3 N2 N1 Desc;
4
4   set pi=3.14159265359;
5
5   smpl 1 21071;
6
6   ?***construção de variáveis para os modelos gama e inversa
    gaussiana***
6   limsup=0.7*caprisc;
7
8   franquia=f;
9   lim=franquia;
10   yy=custo;
11   y0=custo>=limsup; y1=custo<limsup;
13   y=yy*y1+limsup*y0;
14
14   ?***construção de variáveis para o modelo lognormal***
14   lcr=log(caprisc);
15   lcusto=log(custo);
16   llimsup=log(0.7*caprisc);
17   llim=log(franquia);
18   ly0=lcusto>=llimsup; ly1=lcusto<llimsup;
20   ly=lcusto*ly1+llimsup*ly0;
21
21   ?***definição de restrições sobre os parâmetros***
21   n23=n2+n3;
22   fr12=fr1+fr2;
23   ic=ic1+ic2;
24   iv234=iv2+iv3+iv4;
25
25   ?*************************
25   ?modelo Log-normal
25   ?*************************
25
25   title 'Modelo lognormal do custo com sinistros';
26
26   frml eq1
     flv=ly1*((-lcusto-0.5*log(2*pi*sigma^2)-(lcusto-vec)^2/(2*sigma^2))-log(1-
     cnorm((llim-vec)/sigma)))+ly0*(log(1-cnorm((llimsup-vec)/sigma))-log(1-
     cnorm((llim-vec)/sigma)));
27   frml eqvec
     vec=(c0+ccap*lcr+csex1*Sex1+ctc1*TC1+cfr12*Fr12+cfr3*Fr3+ccrc2*CRC2+ccc1*
     CC1+ccc2*CC2+civ1*IV1+civ234*IV234+cic*IC+cac1*AC1+cs3*s3+cs2*s2+cn4*n4+
     cn23*n23+cn1*n1);
28   eqsub(name=mod_lognormal) eq1 eqvec;
29   param c0 7 ccap .4 sigma 1 csex1 ctc1 cfr12 cfr3 ccrc2 civ234 ccc1
     ccc2 civ1 civ2 civ3 civ4 cic cac1 cs3 cs2 cn4 cn23 cn1;
30   ML(hiter=n,hcov=n) mod_lognormal;
31
31   ?*********************
```

```
31  ?modelo Gamma
31  ?*********************
31
31  title 'Modelo gama do custo com sinistros';
32
32  Frml eq1
    flv=y1*((y*(-1/vec)+log(1/vec))*m+log(y*m)*m-log(y)-log(gamfn(m))-log(1-
    cnorm(((9*m)**0.5)*((lim/vec)**(1/3)+1/(9*m)-1))))+y0*(log(1-cnorm(((9*m)**
    0.5)*((limsup/vec)**(1/3)+1/(9*m)-1)))-log(1-cnorm(((9*m)**0.5)*((lim/vec)*
    *(1/3)+1/(9*m)-1))));
33  frml eqvec
    vec=exp(c0+ccap*lcr+csex1*Sex1+ctc1*TC1+cfr12*Fr12+cfr3*Fr3+ccrc2*CRC2+
    ccc1*CC1+ccc2*CC2+civ1*IV1+civ234*IV234+cic*IC+cac1*AC1+cs3*s3+cs2*s2+cn4*
    n4+cn1*n1);
34  eqsub(name=mod_gama) eq1 eqvec;
35  param c0 ccap m .8422 csex1 ctc1 cfr12 cfr3 ccrc2 ccc1 ccc2 civ1
    civ234 cic cac1 cs3 cs2 cn4 cn1;
36  ML(hiter=n,hcov=n) mod_gama;
37
37  ?*********************************
37  ?modelo inversa Gaussiana
37  ?*********************************
37
37  title 'Modelo inversa Gaussiana do custo com sinistros';
38
38  Frml eq1 flv=y1*
    (-fi*y/(2*vec)+fi-vec*fi/(2*y)+0.5*(log(vec)+log(fi)-log(2*pi)-3*log(y))-
    log(1-cnorm((lim/vec-1)*sqrt(fi*vec/lim))-exp(2*fi)*cnorm(-(lim/vec+1)*
    sqrt(fi*vec/lim))))+y0*(log(1-cnorm((limsup/vec-1)*sqrt(fi*vec/limsup))-
    exp(2*fi)*cnorm(-(limsup/vec+1)*sqrt(fi*vec/limsup)))-log(1-cnorm((lim/vec-
    1)*sqrt(fi*vec/lim))-exp(2*fi)*cnorm(-(lim/vec+1)*sqrt(fi*vec/lim))));
39  frml eqvec
    vec=exp(c0+ccap*lcr+csex1*Sex1+ctc1*TC1+cfr12*Fr12+cfr3*Fr3+ccrc2*CRC2+
    ccc1*CC1+ccc2*CC2+civ1*IV1+civ234*IV234+cic*IC+cac1*AC1+cs3*s3+cs2*s2+cn4*
    n4+cn23*n23+cn1*n1);
40  eqsub(name=mod_inv_gaussiana) eq1 eqvec;
41  param c0 ccap fi .8422 csex1 ctc1 cfr12 cfr3 ccrc2 ccc1 ccc2 civ1
    civ234 cic cac1 cs3 cs2 cn4 cn23 cn1;
42  ML(hiter=n,hcov=n) mod_inv_gaussiana;
43
43
```

EXECUTION

```
Current sample:  1 to 21129

Current sample:  1 to 21071
```

Modelo lognormal do custo com sinistros

MAXIMUM LIKELIHOOD ESTIMATION

EQUATION: MOD_LOGNORMAL

CONSTANTS:

	PI
VALUE	3.14159

Working space used: 955757

STARTING VALUES

	C0	CCAP	CSEX1	CTC1	CFR12
VALUE	7.00000	0.40000	0.00000	0.00000	0.00000

	CFR3	CCRC2	CCC1	CCC2	CIV1
VALUE	0.00000	0.00000	0.00000	0.00000	0.00000

	CIV234	CIC	CAC1	CS3	CS2
VALUE	0.00000	0.00000	0.00000	0.00000	0.00000

	CN4	CN23	CN1	SIGMA
VALUE	0.00000	0.00000	0.00000	1.00000

```
F=  0.27560E+06 FNEW=  0.27442E+06 ISQZ=  1 STEP= 0.50000     CRIT=  3755.1
F=  0.27442E+06 FNEW=  0.27433E+06 ISQZ=  0 STEP=  1.0000     CRIT=  159.65
F=  0.27433E+06 FNEW=  0.27433E+06 ISQZ=  0 STEP=  1.0000     CRIT=  5.0554
F=  0.27433E+06 FNEW=  0.27433E+06 ISQZ=  0 STEP=  1.0000     CRIT= 0.96418E-02
F=  0.27433E+06 FNEW=  0.27433E+06 ISQZ=  0 STEP=  1.0000     CRIT= 0.38519E-07

CONVERGENCE ACHIEVED AFTER   5 ITERATIONS

   11 FUNCTION EVALUATIONS.

Number of observations = 21071.0  Log likelihood = -274332.
```

Parameter	Estimate	Standard Error	t-statistic	P-value
C0	7.33869	.172942	42.4344	[.000]
CCAP	.365392	.012449	29.3521	[.000]
CSEX1	.043038	.014714	2.92492	[.003]
CTC1	-.311300	.024730	-12.5881	[.000]
CFR12	.062823	.016615	3.78116	[.000]
CFR3	.135308	.031334	4.31831	[.000]
CCRC2	-.142247	.015447	-9.20855	[.000]
CCC1	.111779	.014850	7.52734	[.000]
CCC2	.258946	.040589	6.37976	[.000]
CIV1	-.091718	.020154	-4.55084	[.000]
CIV234	-.074407	.015882	-4.68492	[.000]
CIC	.159966	.043572	3.67132	[.000]
CAC1	.210955	.056951	3.70412	[.000]
CS3	.115468	.030355	3.80390	[.000]
CS2	.175119	.029728	5.89064	[.000]
CN4	-.068920	.018762	-3.67338	[.000]
CN23	.058604	.022118	2.64966	[.008]
CN1	.145780	.038807	3.75650	[.000]
SIGMA	.884694	.547146E-02	161.692	[.000]

```
Standard Errors computed from   analytic second derivatives
(Newton)
```

Modelo gama do custo com sinistros

MAXIMUM LIKELIHOOD ESTIMATION

```
EQUATION: MOD_GAMA

Working space used: 947751
```

STARTING VALUES

	C0	CCAP	CSEX1	CTC1	CFR12
VALUE	7.33869	0.36539	0.043038	-0.31130	0.062823

	CFR3	CCRC2	CCC1	CCC2	CIV1
VALUE	0.13531	-0.14225	0.11178	0.25895	-0.091718

	CIV234	CIC	CAC1	CS3	CS2
VALUE	-0.074407	0.15997	0.21095	0.11547	0.17512

	CN4	CN1	M

```
VALUE          -0.068920        0.14578        0.84220

F=  0.27604E+06 FNEW=  0.27533E+06 ISQZ=  0 STEP=  1.0000     CRIT=  1304.3
F=  0.27533E+06 FNEW=  0.27523E+06 ISQZ=  0 STEP=  1.0000     CRIT=  222.88
F=  0.27523E+06 FNEW=  0.27523E+06 ISQZ=  0 STEP=  1.0000     CRIT=  11.033
F=  0.27523E+06 FNEW=  0.27523E+06 ISQZ=  0 STEP=  1.0000     CRIT= 0.93546E-02
F=  0.27523E+06 FNEW=  0.27523E+06 ISQZ=  1 STEP= 0.50000     CRIT= 0.12216E-07

CONVERGENCE ACHIEVED AFTER   5 ITERATIONS

   10 FUNCTION EVALUATIONS.

Number of observations = 21071.0  Log likelihood = -275229.

                        Standard
Parameter  Estimate       Error       t-statistic   P-value
C0         5.75798      .176166       32.6850       [.000]
CCAP       .498498      .012657       39.3855       [.000]
CSEX1      .046126      .015622       2.95261       [.003]
CTC1       -.280650     .023776       -11.8039      [.000]
CFR12      .049299      .017622       2.79754       [.005]
CFR3       .125786      .030187       4.16695       [.000]
CCRC2      -.143974     .016616       -8.66500      [.000]
CCC1       .108279      .015610       6.93654       [.000]
CCC2       .256316      .041852       6.12434       [.000]
CIV1       -.124223     .020697       -6.00187      [.000]
CIV234     -.125297     .016701       -7.50216      [.000]
CIC        .130826      .047796       2.73718       [.006]
CAC1       .227804      .063594       3.58218       [.000]
CS3        .104755      .032312       3.24196       [.001]
CS2        .202616      .031708       6.38999       [.000]
CN4        -.111552     .018924       -5.89482      [.000]
CN1        .168831      .041244       4.09345       [.000]
M          1.06539      .017092       62.3329       [.000]

Standard Errors computed from   analytic second derivatives
(Newton)

                Modelo inversa Gaussiana do custo com sinistros
                ===============================================

                        MAXIMUM LIKELIHOOD ESTIMATION
                        =============================

EQUATION: MOD_INV_GAUSSIANA

                                CONSTANTS:

                    PI
VALUE           3.14159

Working space used: 1045259
                                STARTING VALUES

                    C0          CCAP        CSEX1       CTC1        CFR12
VALUE             5.75798       0.49850     0.046126    -0.28065    0.049299

                    CFR3        CCRC2       CCC1        CCC2        CIV1
VALUE             0.12579       -0.14397    0.10828     0.25632     -0.12422

                    CIV234      CIC         CAC1        CS3         CS2
VALUE             -0.12530      0.13083     0.22780     0.10476     0.20262

                    CN4         CN23        CN1         FI
VALUE             -0.11155      0.058604    0.16883     0.84220
```

```
F=  0.27460E+06 FNEW=  0.27431E+06 ISQZ=   0 STEP=  1.0000     CRIT=  707.68
F=  0.27431E+06 FNEW=  0.27429E+06 ISQZ=   0 STEP=  1.0000     CRIT=  30.901
F=  0.27429E+06 FNEW=  0.27429E+06 ISQZ=   0 STEP=  1.0000     CRIT= 0.32485E-01
F=  0.27429E+06 FNEW=  0.27429E+06 ISQZ=   0 STEP=  1.0000     CRIT= 0.51606E-07

CONVERGENCE ACHIEVED AFTER   4 ITERATIONS


    8 FUNCTION EVALUATIONS.

Number of observations = 21071.0  Log likelihood = -274293.

                          Standard
Parameter  Estimate         Error       t-statistic   P-value
C0         7.50437        .167241       44.8715       [.000]
CCAP       .382303        .011956       31.9756       [.000]
CSEX1      .043587        .014123       3.08618       [.002]
CTC1       -.312098       .024031       -12.9875      [.000]
CFR12      .057466        .015964       3.59972       [.000]
CFR3       .126631        .030414       4.16353       [.000]
CCRC2      -.126180       .014797       -8.52721      [.000]
CCC1       .107076        .014311       7.48231       [.000]
CCC2       .249981        .038373       6.51452       [.000]
CIV1       -.085022       .019183       -4.43204      [.000]
CIV234     -.077786       .015213       -5.11316      [.000]
CIC        .154388        .041742       3.69861       [.000]
CAC1       .201129        .053859       3.73434       [.000]
CS3        .114174        .028985       3.93911       [.000]
CS2        .153787        .027848       5.52244       [.000]
CN4        -.063725       .018204       -3.50069      [.000]
CN23       .063478        .021441       2.96052       [.003]
CN1        .137946        .036515       3.77784       [.000]
FI         .943225        .013440       70.1829       [.000]

Standard Errors computed from   analytic second derivatives
(Newton)

********************************************************************************
```

Resultados da modelização dos custos com sinistros com restrição sobre o parâmetro do valor seguro

```
          PROGRAM
LINE  ***************************************************************
1   options crt memory=65;
2   freq n;
3
3   read(file='autozona.txt ') Custo F AS1 AS2 Sex1 Sex2 Cat1 Cat2 Cat3
    Cat4 Cat5 TC1 Fr1 Fr2 Fr3 B1 B2 B3 B4 B5 B6 B7 CRC1 CRC2 PB1 PB2 PB3 CC1
    CC2 IV1 IV2 IV3 IV4 IV5 IV6 IV7 IV8 IV9 IV10 IC1 IC2 AC1 CAPRISC Z1 Z2 Z3
    Z4 Z5 Z6 Z7 Z8 S3 S2 S1 N4 N3 N2 N1 Desc;
4
4   set pi=3.14159265359;
5
5   smpl 1 21071;
6
6   ?***Definição de variáveis***
7   franquia=f;
8   lcusto=log(custo);
9   lcr=log(caprisc);
10   llimsup=log(0.7*caprisc);
11   llimmin=log(franquia);
12   ly0=lcusto>=llimsup; ly1=lcusto<llimsup;
14   ly=lcusto*ly1+llimsup*ly0;
15
15   ?***Definição de restrições sobre os parâmetros***
15   n23=n2+n3;
16   fr12=fr1+fr2;
17   ic=ic1+ic2;
18   iv8910=iv8+iv9+iv10;
19
19   frml eq1
     flv=ly1*((-lcusto-0.5*log(2*pi*sigma^2)-(lcusto-vec)^2/(2*sigma^2))-log(1-
     cnorm((llimmin-vec)/sigma)))+ly0*(log(1-cnorm((llimsup-vec)/sigma))-log(1-
     cnorm((llimmin-vec)/sigma)));
20   frml eqvec
     vec=(c0+ccap*lcr+ctc1*TC1+cfr12*Fr12+cfr3*Fr3+ccrc2*CRC2+ccc1*CC1+ccc2*CC2+
     civ2*IV2+civ3*iv3+civ4*iv4+civ5*iv5+civ6*iv6+civ7*iv7+civ8910*iv8910+cic*
     IC+cac1*AC1+cs3*s3+cs2*s2+cn4*n4+cn23*n23+cn1*n1);
21   eqsub(name=mod_lognormal) eq1 eqvec;
22   set ccap=1;
23   param c0 7.3 sigma 1 ctc1 cfr12 cfr3 ccrc2 ccc1 ccc2 civ2 civ3 civ4
     civ5 civ6 civ7 civ8910 cic cac1 cs3 cs2 cn4 cn23 cn1;
24   ML(hiter=n,hcov=n) mod_lognormal;
25
25
          EXECUTION
*************************************************************************

Current sample:  1 to 21071

                        MAXIMUM LIKELIHOOD ESTIMATION
                        =============================


EQUATION: MOD_LOGNORMAL

                                    CONSTANTS:

                    CCAP                PI
VALUE            1.00000            3.14159

CONVERGENCE ACHIEVED AFTER    5 ITERATIONS

    10 FUNCTION EVALUATIONS.


Number of observations = 21071.0  Log likelihood = -275225.
```

Parameter	Estimate	Standard Error	t-statistic	P-value
C0	-1.97537	.021677	-91.1283	[.000]
CTC1	-.357698	.026301	-13.6001	[.000]
CFR12	.070337	.017776	3.95676	[.000]
CFR3	.109114	.033785	3.22966	[.001]
CCRC2	-.204751	.016437	-12.4566	[.000]
CCC1	-.117960	.015014	-7.85659	[.000]
CCC2	-.202243	.042266	-4.78505	[.000]
CIV2	.093546	.020909	4.47391	[.000]
CIV3	.208220	.023007	9.05022	[.000]
CIV4	.333549	.025888	12.8845	[.000]
CIV5	.501189	.029473	17.0049	[.000]
CIV6	.636787	.032363	19.6765	[.000]
CIV7	.749710	.035981	20.8365	[.000]
CIV8910	.796345	.027120	29.3634	[.000]
CIC	.172742	.046644	3.70338	[.000]
CAC1	.221529	.061005	3.63131	[.000]
CS3	.136791	.032502	4.20870	[.000]
CS2	.167068	.031814	5.25140	[.000]
CN4	-.057404	.020150	-2.84885	[.004]
CN23	.086491	.023730	3.64488	[.000]
CN1	.154891	.041602	3.72313	[.000]
SIGMA	.939149	.599072E-02	156.767	[.000]

Standard Errors computed from analytic second derivatives
(Newton)

****

Resultados das modelizações do custo com sinistros de perda parcial e da probabilidade de perda total (referentes à abordagem bi-etápica)

```
          PROGRAM
LINE  ************************************************************************
1   options crt memory=55;
2   freq n;
3
3   read(file='autozona.txt ') Custo F AS1 AS2 Sex1 Sex2 Cat1 Cat2 Cat3
    Cat4 Cat5 TC1 Fr1 Fr2 Fr3 B1 B2 B3 B4 B5 B6 B7 CRC1 CRC2 PB1 PB2 PB3 CC1
    CC2 IV1 IV2 IV3 IV4 IV5 IV6 IV7 IV8 IV9 IV10 IC1 IC2 AC1 CAPRISC Z1 Z2 Z3
    Z4 Z5 Z6 Z7 Z8 S3 S2 S1 N4 N3 N2 N1 Desc;
4
4   set pi=3.14159265359;
5
5   ?*********************************
5   ?Modelo Custo Perda Total
5   ?*********************************
5
5   title 'Modelo dos Custos de Perdas Parciais';
6
6   smpl 1 19884; ?observações referentes às perdas parciais
7
8   franquia=f;
9   lcr=log(caprisc);
10   llimsup=log(0.7*caprisc);
11   llimmin=log(franquia);
12   lcusto=log(custo);
13
13   ?***definição de restrições sobre os parâmetros***
13   ic=ic1+ic2;
14   fr123=fr1+fr2+fr3;
15
15   frml eq1
     flv=((-lcusto-0.5*log(2*pi*sigma^2)-(lcusto-vec)^2/(2*sigma^2))-log(cnorm((
     llimsup-vec)/sigma)-cnorm((llimmin-vec)/sigma)));
16   frml eqvec
     vec=(c0+ccap*lcr+csex1*Sex1+ctc1*TC1+cfr123*Fr123+ccrc2*CRC2+ccc1*CC1+ccc2*
     CC2+civ1*IV1+cic*IC+cac1*AC1+cs3*s3+cs2*s2+cs1*s1+cn3*n3+cn2*n2+cn1*n1);
17   eqsub(name=c_perda_parcial) eq1 eqvec;
18   param c0 9 ccap .11 sigma .9 csex1 0 ctc1 0 cfr123 0 ccrc2 0 ccc1 0
     ccc2 0 civ1 0 cic 0 cac1 0 cs3 0 cs2 0 cs1 0 cn3 0 cn2 0 cn1 0;
19   ML(hiter=n,hcov=n) c_perda_parcial;
20
20
20
     ?*************************************************************************
     *
20   ?Modelo de Probabilidade de Perda Total
20
     ?*************************************************************************
     *
20
20   read(file='autofrq.txt ') Custo F AS1 AS2 Sex1 Sex2 Cat1 Cat2 Cat3
     Cat4 Cat5 TC1 Fr1 Fr2 Fr3 B1 B2 B3 B4 B5 B6 B7 CRC1 CRC2 PB1 PB2 PB3 CC1
     CC2 IV1 IV2 IV3 IV4 IV5 IV6 IV7 IV8 IV9 IV10 IC1 IC2 AC1 CAPRISC Z1 Z2 Z3
     Z4 Z5 Z6 Z7 Z8 S3 S2 S1 N4 N3 N2 N1 Desc frq1 frq2 frq3 frq4;
21
21   smpl 1 21071; ?observações referentes às perdas parciais e perdas
     totais
22
22   llimsup=log(0.7*caprisc);
23   ptotal=custo>=0.7*caprisc;
24   llim=log(f);
25   lcr=log(caprisc);
26
27   ic=ic1+ic2;
28   fr123=fr1+fr2+fr3;
29
30
31
     vec=(c0+ccap*lcr+csex1*Sex1+ctc1*TC1+cfr123*Fr123+ccrc2*CRC2+ccc1*CC1+ccc2*
     CC2+civ1*IV1+cic*IC+cac1*AC1+cs3*s3+cs2*s2+cs1*s1+cn3*n3+cn2*n2+cn1*n1);
```

```
    ?parâmetro miu da distribuição das perdas parciais para todos os sinistros
 32
 32  pmt=1-pos(1-cnorm((llim-vec)/sigma)/cnorm((llimsup-vec)/sigma));
    ?probabilidade de sinistro ocorrido não ser participado
 33
 34  ?***definição de restrições sobre os parâmetros***
 35  ic=ic1+ic2;
 36  iv23=iv2+iv3;
 37  iv45=iv4+iv5;
 38  n43=n4+n3;
 39
 39  title 'Modelo para valores iniciais dos parâmetros';
 40
 40  frml eq1 flv=ptotal*log(vec/(1-vec))+log(1-vec);
 41  frml eqvec
    vec=(1/(1+exp(-(c0+ccap*lcr+ctc1*TC1+cfr3*Fr3+ccrc2*CRC2+ccc1*CC1+ccc2*CC2+
    civ1*IV1+civ23*IV23+civ45*IV45+civ6*IV6+civ7*IV7+cic*IC+cs2*s2+cn43*n43+
    cn2*n2))));
 42  eqsub(name=p_val_iniciais) eq1 eqvec;
 43  param c0 -3  ccap 0 ctc1 0 cfr3 0 ccrc2 0 ccc1 0 ccc2 0 civ1 0
    civ23 0 civ45 0 civ6 0 civ7 0 cic 0 cs2 0 cn43 0 cn2 0;
 44  ML(hiter=n,hcov=n,silent) p_val_iniciais;
 45
 45  title 'Modelo de probabilidade de Perda Total';
 46
 46  frml eq1 flv=ptotal*log(vec/(1-vec))+log(1-vec);
 47  frml eqvec
    vec=(1/(1+exp(-(c0+ccap*lcr+ctc1*TC1+cfr3*Fr3+ccrc2*CRC2+ccc1*CC1+ccc2*CC2+
    civ1*IV1+civ23*IV23+civ45*IV45+civ6*IV6+civ7*IV7+cic*IC+cs2*s2+cn43*n43+
    cn2*n2))))
 47
    /(1-pmt*(1-1/(1+exp(-(c0+ccap*lcr+ctc1*TC1+cfr3*Fr3+ccrc2*CRC2+ccc1*CC1+
    ccc2*CC2+civ1*IV1+civ23*IV23+civ45*IV45+civ6*IV6+civ7*IV7+cic*IC+cs2*s2+
    cn43*n43+cn2*n2)))));
 48  eqsub(name=p_perda_total) eq1 eqvec;
 49  param c0 ccap ctc1 cfr3 ccrc2 ccc1 ccc2 civ1 civ23 civ45 civ6 civ7
    cic cs2 cn43 cn2;
 50  ML(hiter=n,hcov=n) p_perda_total;
 51
         EXECUTION
*******************************************************************************

Current sample:  1 to 21129

                    Modelo dos Custos de Perdas Parciais
                    ====================================

Current sample:  1 to 19884

                         MAXIMUM LIKELIHOOD ESTIMATION
                         =============================

EQUATION: C_PERDA_PARCIAL

CONVERGENCE ACHIEVED AFTER    9 ITERATIONS

   19 FUNCTION EVALUATIONS.

Number of observations = 19884.0  Log likelihood = -269952.

                          Standard
Parameter  Estimate         Error       t-statistic   P-value
C0         9.46312        .236540       40.0065       [.000]
CCAP       .216185        .016576       13.0422       [.000]
CSEX1      .049021        .017476       2.80496       [.005]
CTC1       -.290326       .022888       -12.6844      [.000]
CFR123     .085814        .017694       4.84980       [.000]
CCRC2      -.171677       .018597       -9.23146      [.000]
CCC1       .132790        .017456       7.60726       [.000]
CCC2       .309611        .046630       6.63978       [.000]
CIV1       -.056180       .019212       -2.92425      [.003]
CIC        .184828        .054016       3.42172       [.001]
CAC1       .241090        .073715       3.27058       [.001]
CS3        .131300        .036378       3.60927       [.000]
CS2        .170188        .035667       4.77154       [.000]
```

CS1	-.092519	.032225	-2.87101	[.004]
CN3	.155203	.031618	4.90873	[.000]
CN2	.102170	.037520	2.72305	[.006]
CN1	.168360	.046578	3.61459	[.000]
SIGMA	.892187	.770128E-02	115.849	[.000]

Standard Errors computed from analytic second derivatives (Newton)

Current sample: 1 to 21071

Modelo para valores iniciais dos parâmetros

Parameter	Estimate	Standard Error	t-statistic	P-value
C0	10.1640	.703770	14.4422	[.000]
CCAP	-.857297	.051650	-16.5983	[.000]
CTC1	-.768737	.143524	-5.35614	[.000]
CFR3	.520262	.174910	2.97445	[.003]
CCRC2	-.270793	.066185	-4.09149	[.000]
CCC1	.220941	.067360	3.27998	[.001]
CCC2	.655430	.186140	3.52117	[.000]
CIV1	-.457119	.105547	-4.33095	[.000]
CIV23	-.761549	.091976	-8.27985	[.000]
CIV45	-.860550	.097626	-8.81472	[.000]
CIV6	-.550489	.122233	-4.50359	[.000]
CIV7	-.323365	.120160	-2.69112	[.007]
CIC	.361572	.160777	2.24891	[.025]
CS2	.274392	.116567	2.35393	[.019]
CN43	-.698542	.086748	-8.05256	[.000]
CN2	-.334963	.152986	-2.18950	[.029]

Standard Errors computed from analytic second derivatives (Newton)

Modelo de probabilidade de Perda Total

MAXIMUM LIKELIHOOD ESTIMATION

EQUATION: P_PERDA_TOTAL

CONVERGENCE ACHIEVED AFTER 3 ITERATIONS

6 FUNCTION EVALUATIONS.

Number of observations = 21071.0 Log likelihood = -4127.86

Parameter	Estimate	Standard Error	t-statistic	P-value
C0	10.5849	.701501	15.0889	[.000]
CCAP	-.889947	.051509	-17.2774	[.000]
CTC1	-.873344	.144118	-6.05994	[.000]
CFR3	.456175	.175593	2.59791	[.009]
CCRC2	-.288212	.066447	-4.33750	[.000]
CCC1	.213545	.067622	3.15794	[.002]
CCC2	.642813	.186318	3.45009	[.001]
CIV1	-.461003	.105932	-4.35186	[.000]
CIV23	-.754093	.092307	-8.16939	[.000]
CIV45	-.846052	.097818	-8.64928	[.000]
CIV6	-.526480	.122437	-4.30000	[.000]
CIV7	-.316321	.120339	-2.62858	[.009]
CIC	.372309	.160933	2.31344	[.021]
CS2	.289097	.116701	2.47725	[.013]
CN43	-.700981	.086943	-8.06254	[.000]
CN2	-.316830	.153073	-2.06980	[.038]

Standard Errors computed from analytic second derivatives (Newton)

****

Resultados da modelização do custo com sinistros de perda total (referente à abordagem bi-etápica)

```
          PROGRAM
LINE  ********************************************************************
 1  options crt;
 2
 2  read(file="SinGran2.xls");
 3
 3  set pi=3.14159265359;
 4
 4  z=(y)/1000000;
 5  liminf=(0.7*caprisc)/1000000;
 6  limsup=(1.3*caprisc)/1000000;
 7  cap=caprisc/1000000;
 8
 8  frml eq1
    flv=-0.5*log(2*pi*(cv*vec)^2)-0.5*(z/(cv*vec)-1/cv)^2-log(cnorm((limsup/(cv*
    vec)-1/cv))-cnorm((liminf/(cv*vec)-1/cv)));
 9  frml eqvec vec=(c0+ccap*cap);
 10  eqsub(name=flvnormal) eq1 eqvec;
 11  param c0 0 ccap 0.7 cv .2;
 12  ML(hiter=n,hcov=n) flvnormal;
 13
 13
 13
          EXECUTION
********************************************************************************

Current sample:  1 to 1187

                          MAXIMUM LIKELIHOOD ESTIMATION
                          =============================

EQUATION: FLVNORMAL

                                    CONSTANTS:

                    PI
VALUE          3.14159

Working space used: 11675
                                  STARTING VALUES

                     C0           CCAP           CV
VALUE           0.00000        0.70000        0.20000

F=  -641.47     FNEW=  -796.38     ISQZ=  0 STEP=  1.0000     CRIT=  286.16
F=  -796.38     FNEW=  -815.80     ISQZ=  0 STEP=  1.0000     CRIT=  31.448
F=  -815.80     FNEW=  -821.49     ISQZ=  0 STEP=  1.0000     CRIT=  9.1950
F=  -821.49     FNEW=  -823.02     ISQZ=  0 STEP=  1.0000     CRIT=  2.5406
F=  -823.02     FNEW=  -823.29     ISQZ=  0 STEP=  1.0000     CRIT= 0.47229
F=  -823.29     FNEW=  -823.31     ISQZ=  0 STEP=  1.0000     CRIT= 0.33414E-01
F=  -823.31     FNEW=  -823.31     ISQZ=  0 STEP=  1.0000     CRIT= 0.27270E-03
F=  -823.31     FNEW=  -823.31     ISQZ=  0 STEP=  1.0000     CRIT= 0.21633E-07

CONVERGENCE ACHIEVED AFTER   8 ITERATIONS

   16 FUNCTION EVALUATIONS.

Number of observations = 1187.00  Log likelihood = 823.311

                        Standard
Parameter  Estimate       Error       t-statistic   P-value
C0         .029991     .682199E-02    4.39618       [.000]
CCAP       .647341     .035285        18.3460       [.000]
CV         .299998     .035563        8.43577       [.000]

Standard Errors computed from   analytic second derivatives
(Newton)
********************************************************************************
```